2020

Reconstruction of Bacterial Strain Genomes from Shotgun Metagenomic Reads

Xin Li
University of Central Florida

RECONSTRUCTION OF BACTERIAL STRAIN GENOMES FROM SHOTGUN METAGENOMIC READS

by

XIN LI

B.E. Yanshan University, 2010

M.S. Florida International University, 2013

A dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Computer Science

in the Department of Computer Science

in the College of Engineering and Computer Science

at the University of Central Florida

Orlando, Florida

Fall Term

2020

Major Professor: Haiyan Hu

ABSTRACT

It is necessary to study bacterial strains in environmental samples. The environmental samples are mixed DNA samples collected from the ocean, soil, lake, human body sites, etc. In a natural environment, they provide us new insights into the diversity of our earth. As for bacterial strains on or inside human bodies, to select the proper treatment for diseases caused by bacterial strains, it is critical to identify the corresponding strains and reconstruct their genomes. However, it is a challenge to do so with the DNA from a large number of unknown microbial species mixed together in an environmental sample. The majority of available computational methods depend on available sequenced genomes and marker genes, which can not fully discover the strains and reconstruct their genomes from the shotgun metagenomic reads.

In this dissertation, we studied bacterial strain reconstruction, including one case study about shotgun metagenomic sequencing and two novel approaches to improve the performance of reconstructing bacterial strains. Firstly, we studied how newly sequenced genomes affect the analysis result from shotgun metagenomic datasets. In this study, we found two more new phyla that were related to colitis development compared with a previous study, and the two new phyla were also more statistically significant. Furthermore, we found that one major conclusion from the previous study was not supported by repeating the analysis with an updated marker gene database and tools in metagenomics. Secondly, to better analyze shotgun metagenomic datasets, BHap, a novel algorithm based on fuzzy flow networks and de Bruijn graph was developed to reconstruct bacterial strains. BHap had high precision, recall and F1 score and low susceptibility to sequence errors. It also outperformed existing tools in terms of better precision, better recall, higher F1 score and more accurate estimation of the number of strains. Last but not least, a second approach, mixtureS, was developed by considering all genome positions. MixtureS is based on the EM

algorithms and the frequency difference of strains to distinguish different strains of a bacterial species in shotgun metagenomic datasets. Compared with several existing methods including BHap, mixtureS had a better performance in terms of precision, recall, the prediction accuracy of the strain numbers and abundance. Based on the developed BHap and mixtureS methods, we also developed two software tools, which will be valuable for future strain studies in metagenomics.

ACKNOWLEDGMENTS

I would sincerely express my gratitude to my advisor Dr. Haiyan Hu, who continually teaches me the importance of a good attitude for research and work. That is extremely important for my Ph.D. studies. It would be not possible to finish this dissertation without her guidance and help.

I would also like to thank my co-advisor Dr.Xiaoman Li, who is a talented scientist. I have learned a lot from him for the past six years. He has given countless thoughtful suggestions to my research, and guide me in the right research direction.

I also want to this opportunity to thank my committee members Dr. Saleh A. Naser and Dr. Liqiang Wang for serving in my committee. Their suggestions are very helpful for me to complete my dissertation.

Finally, I would like to thank my parents. They are supporting me and encouraging me in many different ways through my whole graduate study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

1.1 Metagenomics

Metagenomics has been widely defined as analysis of DNA collecting directly from the environment (Hugenholtz & Tyson, 2008). Environmental samples are mixed DNA samples collected from soil, ocean, lake, acid mine drainage and human gut, etc. (Breitbart et al., 2002; Hugenholtz, 2002; Tyson et al., 2004). The estimated total number of microbes is 10^{30} (Turnbaugh & Gordon, 2008). The ocean is home for most microbes, and the estimation of single cell organisms can be reached to $2.9 * 10^{29}$ in 2012 (Kallmeyer, Pockalny, Adhikari, Smith, & D'Hondt, 2012; Lougheed, 2012). Researchers have found more than 5000 different viruses from seawater in 2002 (Breitbart et al., 2002). Bacteria and archaea have also been found in environmental samples from early 16S rRNA sequences, and those 16s RNA sequences are not categorized into any known cultured species (Hugenholtz, Goebel, & Pace, 1998). Bacteria, archaea, and microeukaryotes are essential to all kinds of life, because they are the major source of nutrients, and the key recyclers that can change the dead matter back into available organic form (Bäckhed, Ley, Sonnenburg, Peterson, & Gordon, 2005; Hooper, Midtvedt, & Gordon, 2002; Wooley, Godzik, & Friedberg, 2010).

Microbes also have both positive and negative effects on human condition from human body, farm animal, agriculture, food industry and medicine development (Bäckhed et al., 2005; Berg, 1996; Hooper et al., 2002; Savage, 1977). Understanding human condition is an essential key to understand human genome (Collins & McKusick, 2001; Kaput et al., 2009).More than 100 trillion

microbial cells reside in humans that is around 10-fold more than human cells, and those microbial cells can encode genes 100 times larger than human cells (Ley, Peterson, & Gordon, 2006). In terms of advantage, microbes are offering many benefits to the humans. It can affect the host by physiological functions like regulating host immunity (Gensollen, Iyer, Kasper, & Blumberg, 2016), providing protection to host from pathogens (Bäumler & Sperandio, 2016), exerting several beneficial effects on energy metabolism (Den Besten et al., 2013). Besides of benefits, microbes are also closely related to some intestinal disease and distant organs disease such as obesity, inflammatory diseases and colorectal cancer (Chang & Lin, 2016; Gueimonde, Ouwehand, Huhtinen, Salminen, & Salminen, 2007; Schroeder & Bäckhed, 2016). Although there are large diversities on species level, majority of the gut microbial environment is composed of five phyla: Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria and Verrucomicrobia (Schroeder & Bäckhed, 2016).

In the late 1970s, researchers studied in sequencing microbial genome bacteriophages MS2 and φ-X174 (Fiers et al., 1976; Sanger et al., 1977). The first sequenced bacterial genome is Haemophilus influenza with 1.8 million base pairs in 1995, which shed lights on complete genome sequence from a free-living organism (Fleischmann et al., 1995). The next major step about metagenomics is the Global Ocean Sampling Expedition (GOS), which is to study the diversity in marine microbial environmental samples. The researchers spent two year exploring the West Coast of the United States, Baltic, Mediterranean and Black Seas (Rusch et al., 2007; Venter et al., 2004; Yooseph et al., 2007). Through their journey to Sargasso Sea, around 2000 different species were found by DNA analysis and 148 of them were unknown bacteria (Venter et al., 2004). By the time of this dissertation, there are total 1801, 28475, 20545 different types of genomes that have one or

more many genome sequencing projects that may be complete, in progress or planned for Archaea, Bacteria and Virus respectively (ftp://ftp.ncbi.nlm.nih.gov/genomes/).

Although such a large number of single genomes has been sequenced, the limitation of a single genome can not be neglected. Firstly, if researchers want to sequence a single microbe, the microbe has to be cloned before sequencing. However, the limitation is that only a small percentage of environmental microbes can be cloned in lab, which will cause a high bias in cloned samples that can not be used to draw the full picture of environmental community (Amann, Ludwig, & Schleifer, 1995; Pace, 1997; Rappé & Giovannoni, 2003). Secondly, in nature, it rarely sees that an environmental sample only consists of one species (Wooley et al., 2010). Most species may have interaction with other species, such as species in the human gut. By these two reasons, a traditional clone method will fail to represent the full picture of the environmental community. However, the new sequencing technology with reducing cost of sequencing can overcome such limitations, like 16S ribosomal RNA and shotgun sequencing. For those new sequencing technology, a large number of species can be sequencing directly from environmental samples instead of sequencing individual species (Handelsman, Rondon, Brady, Clardy, & Goodman, 1998; Rondon et al., 2000). The next big impact on metagenomics research is from next-generation sequencing platforms. Platforms like the Genome Analyzer of Illumina , the SOLiD system of Applied Biosystems (Bentley, 2006) and the Roche 454 sequencer (Margulies et al., 2005) is in the advantage of cost reduction and throughput so that a largely increasing number of metagenomic project has been emerged like Global Ocean Sampling project (Rusch et al., 2007; Venter et al., 2004; Yooseph et al., 2007), MetaHit and the Human Microbiome Project (Hugon et al., 2015; J. Li et al., 2014).

1.2 16S ribosomal RNA and shotgun sequencing

Previously, culturable bacteria and archaea can be identified by agar media-based methods or biochemical tests. With the advent of next generation sequencing technologies (Muegge et al., 2011; Qin et al., 2010), 16s rRNA and shotgun sequence can be used for identifying both culturable and unculturable bacteria and archaea. In early study, 16s ribosomal RNA (rRNA) is a major sequencing technology in metagenomics samples because16S rRNA gene was conserved in most bacterial and archaea and there are also nine highly variable regions (Thursby & Juge, 2017). By such property, species with 16S rRNA genes can be easily identified. In the early stage of 16s rRNA, sequencing the whole gene is the routine procedure. However, because of high bias and insensitivity, researchers are changing to a shorter region but with depth sequencing (Suau et al., 1999).

Although many experiments are based on 16s rRNA, several biases have been found by replicating the experiments from the same biological sample. It also has limited resolution and lower sensitivity problems (Poretsky, Rodriguez-R, Luo, Tsementzi, & Konstantinidis, 2014). The whole-genome shotgun sequencing technology has become popular with more reliable estimates of diversity and composition of the microbial world, which is caused by higher sensitivity and resolution (Poretsky et al., 2014). Another reason is that shotgun sequencing can avoid bias from amplification of phylogenetic marker genes (Simon & Daniel, 2011). Two major publications for shotgun sequencing applying on metagenomics are assessing the diversity and composition from mine drainage biofilm (Tyson et al., 2004) and the Sargasso Sea (Venter et al., 2004).

Although we can gain benefit from new sequencing technology, the difficulty in assembling the environmental sample is also large because of fragmented sequences. Since fragments could be from different species or same species with different strains, it increases the difficulty to find the true species or strains in an environmental sample. Under such circumstances, new algorithms in

bioinformatics are necessary in metagenomic research. The process of different species is clustered into individual species or Operational Taxonomic Units(OTUs) is called read binning (Eisen, 2007). The binning methods can be further divided into taxonomy-dependent methods and taxonomy-independent methods. Taxonomy-dependent methods depend on the known reference genome by checking the similarity of reads, k-mers or compositional properties such as GC content. The reads binning methods by checking similarity of reads include MBMC (Ying Wang, Hu, & Li, 2016), MEGAN (Huson, Auch, Qi, & Schuster, 2007), MLTreeMap(Stark, Berger, Stamatakis, & von Mering, 2010) and SOrt-ITEMs (Monzoorul Haque, Ghosh, Komanduri, & Mande, 2009). Kraken assigns reads on taxonomical trees by counting k-mers in reference genome (Wood & Salzberg, 2014). Methods based on compositional properties take GC content, oligonucleotide usage patterns or pre-computed models into consideration, such as NBC (Rosen, Reichenberger, & Rosenfeld, 2011) and Phymm (Brady & Salzberg, 2009). In contrast, Taxonomy-independent methods do not consider the reference genome, and they will bin reads by comparing the difference of GC content, k-mer frequencies from different species in a environmental sample instead (Ying Wang, 2016). Taxonomy-independent methods include MBBC (Ying Wang, Hu, & Li, 2015), compostBin (Chatterji, Yamazaki, Bai, & Eisen, 2008), AbundanceBin (Wu & Ye, 2011) and MetaCluster (Yi Wang, Leung, Yiu, & Chin, 2012).

1.3 Bacterial strains

Different binning methods can help people to measure the diversity by clustering different species into individual species within a metagenomic sample, but it can not be applied to distinguish bacteria in strain level. Species is a distinct group of strains with some distinguishing features, and a strain is a genetic variant within a species (Brenner, Staley, & Krieg, 2005). In bacteriology, a

strain is also defined as the basic operational unit (Dijkshoorn, Ursing, & Ursing, 2000). In bacterial strains, most phenotype variability can be explained by genetic diversity, like geographic distribution, host specificity, pathogenicity, antibiotic resistance, and virulence (W. Li, Raoult, & Fournier, 2009). In term of human health, bacterial strains also play a vital role, such as antibiotic resistance, increased virulence and transmissibility, host spectra expandability, bioterrorist attacks (Fournier, Zhu, Ogata, & Raoult, 2004; W. Li et al., 2009).

In the early molecular biology, a relative homogeneous population with 70% hybridization of DNA similarity and at least 97% 16s rRNA gene sequence similarity is defined as a bacterial species (W. Li et al., 2009; Stackebrandt & GOEBEL, 1994; L. Wayne et al., 1987; L. G. Wayne, 1988). However, with the increasing number of complete bacterial genomes, genetic diversity is actually much larger than previous study by comparative genome study, and the gene content difference between two strains of one bacterial species can be up to 30%. (Binnewies et al., 2006; Fraser-Liggett, 2005; Lefébure & Stanhope, 2007; Tettelin et al., 2005). Considering the average length of bacteria, the possibility of diversity may be countless. Several genetic forces can lead to strain diversity within bacteria, such as point mutation, genome reduction, genome rearrangement, gene duplication and gene acquisition (Bryant, Chewapreecha, & Bentley, 2012; Darmon & Leach, 2014; Fraser-Liggett, 2005; Levin & Bergstrom, 2000). A study from eight strains of Streptococcus agalactiae has shown that the bacteria genome may be divided into three parts: one core genome shared by all strains, a group of distributed genes shared by some of strains and a group of genes that is specific to some strains (Tettelin et al., 2005). Such bacterial genome division introduces the concept of bacterial pan-genome which include core genome and dispensable genome (Fraser-Liggett, 2005; W. Li et al., 2009; Tettelin et al., 2005).

1.4 Other strain reconstruction method

There are several approaches available for viral strains. ShoRAH firstly estimated the strain diversity in local level, and then reconstructed genome-wide strains by path cover algorithm (Zagordi, Bhattacharya, Eriksson, & Beerenwinkel, 2011). QuRe is to reconstruct viral strains by sliding windows. By sliding windows, it partitions the reference genome and count reads within sliding windows. A score will be given to the partition, and then it will construct an overlap graph. Then a heuristic algorithm will be applied to find a path from the overlap graph (Prosperi & Salemi, 2012). Although the above two methods can be used for reconstructing viral strains, it becomes difficult for them to perform the same task on bacteria. The reason is that the bacterial mutation rate is much lower than viral mutation rate. Such relatively high viral mutation rate can usually make the distance between polymorphic sites shorter than read length. Such information from overlapped reads can be easily applied to reconstruct viral strain. However, the distance between polymorphic sites in bacteria is often longer than several thousand bps (Pulido-Tamayo et al., 2015). Such overlapped reads information can not be applied for bacteria because of no polymorphic reads for many reads.

EVORhA is the first strain reconstruction tool for bacterial population (Pulido-Tamayo et al., 2015). By aligning short reads, EVORhA firstly infer template local strains for each self-defined local window. Secondly, shared polymorphic sites will be used to concatenate templates in the procedure of extending local windows. Lastly, the final genome-wide strain is reconstructed by relative coverage of extended strain from previous procedure. Strain Finder if another tool to infer strain genotypes and track them over time (Smillie et al., 2018). It will tabulate the SNPs for each genome position by aligning all reads against a reference genome. Then a multinomial distribution is used to estimate the strains from alignment data of each given position. Expectation-

maximization (EM) algorithm is then used to find the maximum likelihood estimates of the genotypes and the strain frequency. Exhaustively searching genotypes is performed for estimating strain genotypes.

1.5 Overview of the Dissertation

In summary, we studied the bacterial strain reconstruction. Having a sense of bacterial proportion and genotypes is critical for selecting proper treatment for bacterial caused disease. Reconstructing bacterial strain is the method to solve such problems.

In Chapter 2, I will start with a case study that newly sequenced genomes or methods can affect previous results.

In Chapter 3, I will present a new approach BHap based on fuzzy flow networks and De Bruijn graph, which can achieve high precision, recall and F1 score.

In Chapter 4, I will present a second method mixtureS based on all genome positions and EM algorithms that help this method achieve the state of the art.

CHAPTER TWO: OLD METAGENOMIC DATA MEET NEWLY SEQUENCED GENOMES

Previously published as Li, X., Naser, S. A., Khaled, A., Hu, H., & Li, X. (2018). When old metagenomic data meet newly sequenced genomes, a case study. PloS one, 13(6), e0198773.

2.1 Introduction

A plethora of metagenomic datasets have been generated in the past fifteen years (Breitbart et al., 2002; Poinar et al., 2006; Turnbaugh et al., 2007; Venter et al., 2004). Early datasets are often based on 16S rRNA profiling and Sanger sequencing (Connon & Giovannoni, 2002; Gill et al., 2006; Morris et al., 2002). Later datasets are usually sequenced by next generation sequencing technologies (Muegge et al., 2011; Qin et al., 2010). The generated datasets vary from the early ones such as those in seawater (Breitbart et al., 2002), acid mine drainage (Tyson et al., 2004) and deep sea (Huber et al., 2007; Nakagawa et al., 2004) to current ones such as those in gut (Greenblum, Turnbaugh, & Borenstein, 2012; Qin et al., 2010), skin (Oh et al., 2014), soil (Fierer et al., 2012), etc. These metagenomic datasets have enabled an unprecedented exploration of microbes, which has significantly advanced our understanding of microbes in the living world (Poinar et al., 2006; Qin et al., 2010; Venter et al., 2004). Correspondingly, dozens of computational methods are developed for the analyses of metagenomic datasets. These include methods for filtering erroneous and duplicated reads, methods for gene prediction directly from metagenomic reads, similarity-based and abundance-based methods for read binning, methods for contig binning and genome assembly, etc. (Alneberg et al., 2014; Brady & Salzberg, 2009; Franzosa et al., 2015; Huson et al., 2007; Kim, Song, Breitwieser, & Salzberg, 2016; Krause et al.,

2008; Leung et al., 2011; Markowitz et al., 2007; McHardy, Martín, Tsirigos, Hugenholtz, & Rigoutsos, 2007; Rho, Tang, & Ye, 2010; Segata, Börnigen, Morgan, & Huttenhower, 2013; Segata et al., 2012; Ying Wang et al., 2015, 2016; Ying Wang, Hu, & Li, 2017). These methods altogether have significantly advanced our understanding of the genetic contents in various metagenomic datasets (Handelsman et al., 1998; Kunin, Copeland, Lapidus, Mavromatis, & Hugenholtz, 2008; Wooley et al., 2010).

The majority of available computational methods that perform on metagenomic datasets somewhat rely upon available sequenced genomes. For instance, most methods predict species present in metagenomic datasets depend on the annotation of the available sequenced genomes, such as Megan and MetaPhlAn (Huson et al., 2007; Segata et al., 2012). Megan is an early method that infers species presence based on the comparison of shotgun metagenomic reads with annotated sequences (Huson et al., 2007). MetaPhlAn is a popular method for inferring species present in a metagenomic dataset with marker genes, which infers marker genes from sequenced genomes (Segata et al., 2012). It is understandable that most methods are based on annotated genomes, as more information is taken into account in the analyses and thus more reliable conclusions may be made. Moreover, although metagenomic reads can be studied and analyzed without sequenced genomes, such as read binning and gene prediction, the automatic inference of the origin of a sequence and the presence of a species without any prior information is still infeasible.

Dubin et al. generated a metagenomic dataset to study colitis development in metastatic melanoma patients followed by CTLA4-blockage (Dubin et al., 2016). In their study, shotgun metagenomic reads are sequenced from faecal samples of each of twelve colitis-free (CF) patients and each of ten progressed to colitis (PtC) patients, together with 16S rRNA reads sequenced from faecal samples of each of 34 patients. These 22 patients from whom the shotgun metagenomic reads came

are included in the 34 patients used for 16S rRNA sequencing. This study pointed out that taxonomical analysis results based on 16S rRNA reads from the 34 samples were similar to those based on shotgun metagenomic reads from the 22 samples by the popular method MetaPhlAn (Dubin et al., 2016). In brief, the phylum Bacteroidetes and its three families, Bacteroidaceae, Rikenellaceae, and Barnesiellaceae, were identified to be significantly enriched in CF samples compared with PtC samples (Mann-Whitney test p-value 0.013 for Bacteroidetes, p-value 0.007, 0.023 and 0.013 for the three families, respectively). For simplicity's sake, we used ªbetween samplesº to refer to ªbetween CF samples and PtC samplesº in the following. Moreover, the abundance of reads from Bacteroidetes and its three families negatively correlates with the severity of colitis, with the Spearman's rank correlation coefficient around -0.38, -0.43, -0.42 and -0.43, respectively.

Since this original study was published two years ago (Dubin et al., 2016), genomes of more microbes have been sequenced. Moreover, MetaPhlAn, the tool used in this study, is based on marker genes, which cannot fully utilize the information buried in metagenomic reads (Segata et al., 2012). We thus re-analyzed all shotgun metagenomic reads generated from the 22 patient samples in this metagenomic dataset by mapping reads to all sequenced microbial genomes instead of considering only reads from marker genes (Methods). We considered shotgun metagenomic reads only, as they are more unbiased for taxonomical analysis than 16S rRNA reads (Jovel et al., 2016; Manichanh et al., 2008). Unexpectedly, we found that reads from Bacteroidetes are only marginally more in CF samples than in PtC samples. Moreover, significantly more reads from at least two new phyla, Thaumarchaeota and Actinobacteria, are in PtC samples than in CF samples. The abundance of reads from these two new phyla correlates with the severity of colitis much better than that from Bacteroidetes. By further studying low level taxa based on different strategies,

we found that the read abundances of at least 2 classes, 9 orders, 22 families, 70 genera, and 162 species are significantly different between the two types of samples, and correlate with the severity of colitis in patients better than that of Bacteroidetes. Surprisingly, by repeating the analysis performed in the original study on this dataset with both old and current versions of MetaPhlAn (Dubin et al., 2016), we found that the previously identified phylum Bacteroidetes is not significantly different between samples while one of the newly identified phyla, Actinobacteria, is identified as the only significant phylum between the samples. Our study demonstrated the necessity to reanalyze the generated metagenomic data, the limitation of the marker gene based methods, and the importance of being cautious about the inference from available sequenced genomes in metagenomic studies.

2.2 Materials and Methods

2.2.1 Data and their processing

Pair-end raw read datasets from ten PtC samples and twelve CF samples were downloaded under the BioProject ID: PRJNA302832. There were 78 files in this dataset, in which only 44 files correspond to shotgun metagenomic reads of the 22 patients. We thus only analyzed shotgun metagenomic reads from these 44 files. The program fastq-dump was used to convert raw read datasets into fastq format. Cutadapt was used to cut common adapters and primers with the command: cutadapt -minimum-length 36 -q 3, 3 -a file: common_adapter. After removing adapters and primers, there were still based pairs at the start and the end of reads with low quality or excessively k-mer content based on fastQC. These base pairs were cut by cutadapt with the

command: cutadapt -minimum-length 36 -q 3, 3 -cut 10 -cut -10 -U 10 -U -10. Finally, seqtk was used to convert fastq to fasta (Figure 1).

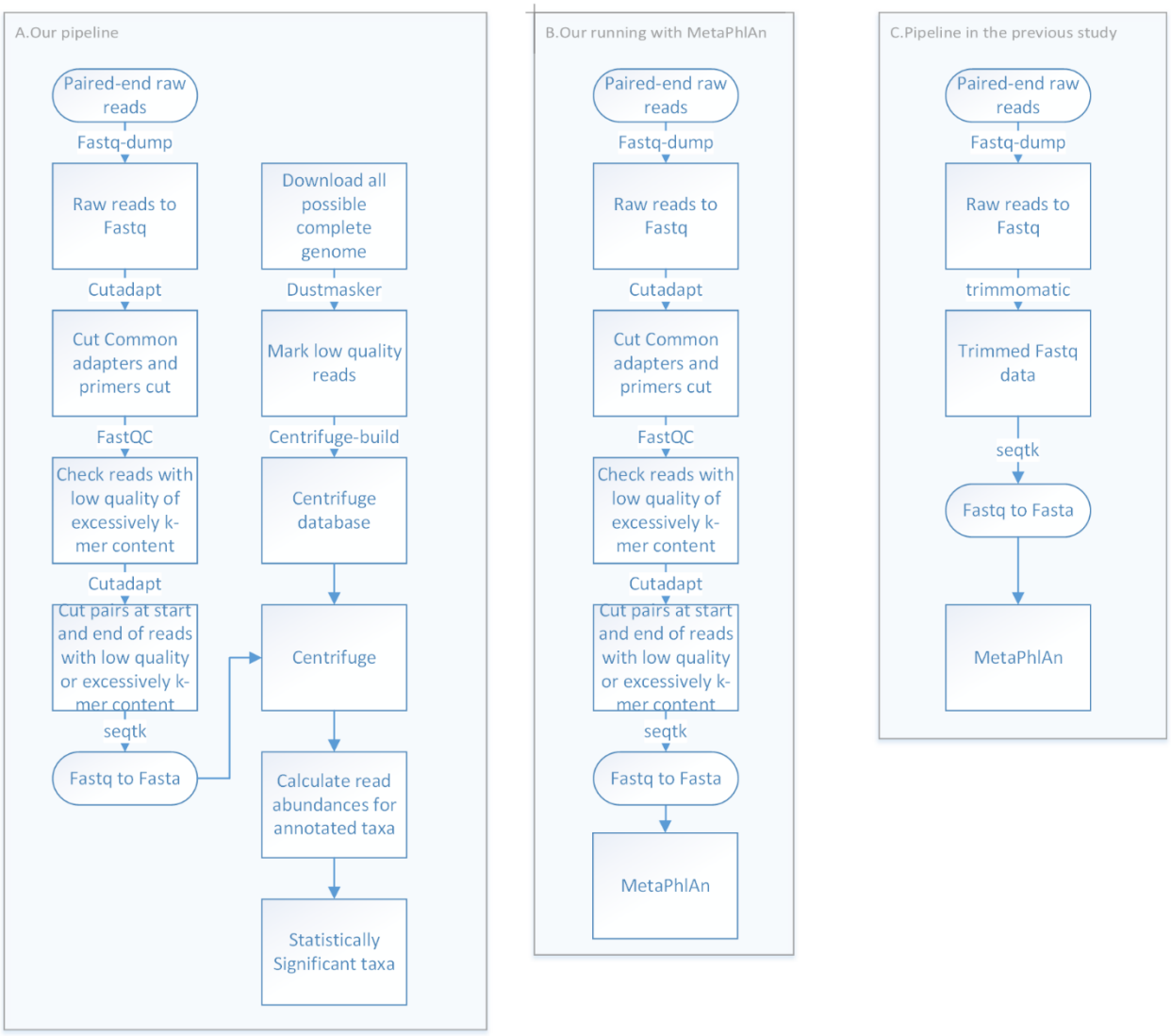

Figure 1: Pipelines to analyze shotgun metagenomic reads

The left panel shows our pipeline. Centrifuge outputs the reads mapped to the sequenced genomes, from which the read abundances of taxa, Mann-Whitney p-values, and correlations with colitis severity are calculated. The middle panel shows the analyses with MetaPhlAn by a different read trimming procedure from that in the original study. The right panel shows the pipeline using MetaPhlAn in the original study. As it is not clear which MetaPhlAn version was used in the original study, two versions of MetaPhlAn have been used for comparisons in this study.

2.2.2 Database preparation and Centrifuge

We mapped the processed reads to sequenced microbial genomes with the Centrifuge tool (Kim et al., 2016). Firstly, all possible complete genomes of archaea, bacteria and viruses were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/), which include 245, 7410 and 7281 complete genome sequences for archaea, bacteria and viruses, respectively. With the corresponding assembly summary file, we found the taxonomy ID of each complete genome sequence. With the detail taxonomy ID file at ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/, we calculated the full lineage information for each complete genome sequence. These sequenced genomes were from 38 phyla, 71 classes, 162 orders, 451 families, 1638 genera, and 9980 species. Secondly, dustmasker was used to mark the low quality reads with command: dustmasker -infmt fasta -in inName -level 20 -outfmt fasta | sed '/^>/! s/[^AGCT]/N/g' > resName. Thirdly, centrifuge- build was used to build the index for Centrifuge with the command: centrifuge-build -p 8 -conversion-table seqid2taxid.map -taxonomy-tree nodes.dmp -name-table names.dmp input-sequences.fna abv. Finally, Centrifuge was used to annotate raw reads with the command: centrifuge -f -p 8 -t -x index_name -1 forward.fasta -2 reverse.fasta -S result -reportfile result_report. Centrifuge gives score(s) to each mapped read. Reads with more than one score are multi-reads that can be mapped to several genomes (Figure 1).

For a sequenced genome, we counted the mapped reads in each sample and normalized this number by dividing the count by the total number of reads in the corresponding sample. We then compared the twelve normalized numbers from the CF samples with the ten normalized numbers from the PtC samples for this sequenced genome. For a taxon at the level higher than the species level, reads from all species contained in this taxon were counted, normalized, and compared.

2.2.3 MetaPhlAn

We inferred the present species and the read abundances by MetaPhlAn with its default parameter (Segata et al., 2012). Since it was not clear which version of MetaPhlAn was used in the original study (Dubin et al., 2016), we applied the two latest versions (version 1.7.7 and version 2.1.0) of MetaPhlAn to the shotgun metagenomic dataset. The input for MetaPhlAn was the same as Centrifuge, which were raw reads in fasta format (Figure 1). The output from MetaPhlAn was the read abundances for each taxon at each taxonomical level. The sum of the abundance of all taxa under the same level was 100%. Then we fetched the abundances for each taxon in each sample for further analyses.

2.2.4 Statistical analysis

Mann-Whitney p-values was calculated with R package, in which the two-sided exact p-value with correction was calculated. The correlation of the read abundances with the severity of colitis was calculate by the Spearman-Rank correlation with python2.7 in the scipy.stats package. The severity of colitis scores were obtained from the original study (Dubin et al., 2016).

2.3 Results

2.3.1 At least two new phyla may relate to colitis development in patients

We mapped shotgun metagenomic reads from each of the 22 faecal samples to about 15,000 sequenced microbial genomes and compared the relative abundance of reads from every phylum in CF samples with that in PtC samples (Figure 1 and Methods). We discovered that the abundance of reads from seven phyla are significantly different between CF samples and PtC samples (Mann-

Whitney p-value < = 0.05), including Bacteroidetes identified previously (Dubin et al., 2016). Five phyla were identified when only uniquely mapped reads were considered. A different set of five phyla were identified when both unique and multi-mapped reads were considered (Figure 2A). Multi-mapped reads are reads that can be mapped to multiple sequenced microbial genomes. For convenience, we call multi-mapped reads and uniquely mapped reads multireads and unique reads, respectively.

With unique reads, we identified five phyla that are significantly different between samples. They are Thaumarchaeota (p-value = 0.009), Actinobacteria (p-value = 0.011), Dictyoglomi (p-value = 0.043), Elusimicrobia (p-value = 0.043), and Bacteroidetes (p-value = 0.050) (Fig 2A). Although Bacteroidetes discovered in the original study is identified, it has the largest Mann-Whitney test p-value, suggesting that the four new phyla are even more significant and may be more related to colitis development. In fact, Bacteroidetes has a negative correlation of -0.335 with the severity of colitis, while Thaumarchaeota, Actinobacteria, Elusimicrobia, and Dictyoglomi have a similar or higher positive correlation of 0.504, 0.480, 0.358, and 0.322, respectively (Figure 2B).

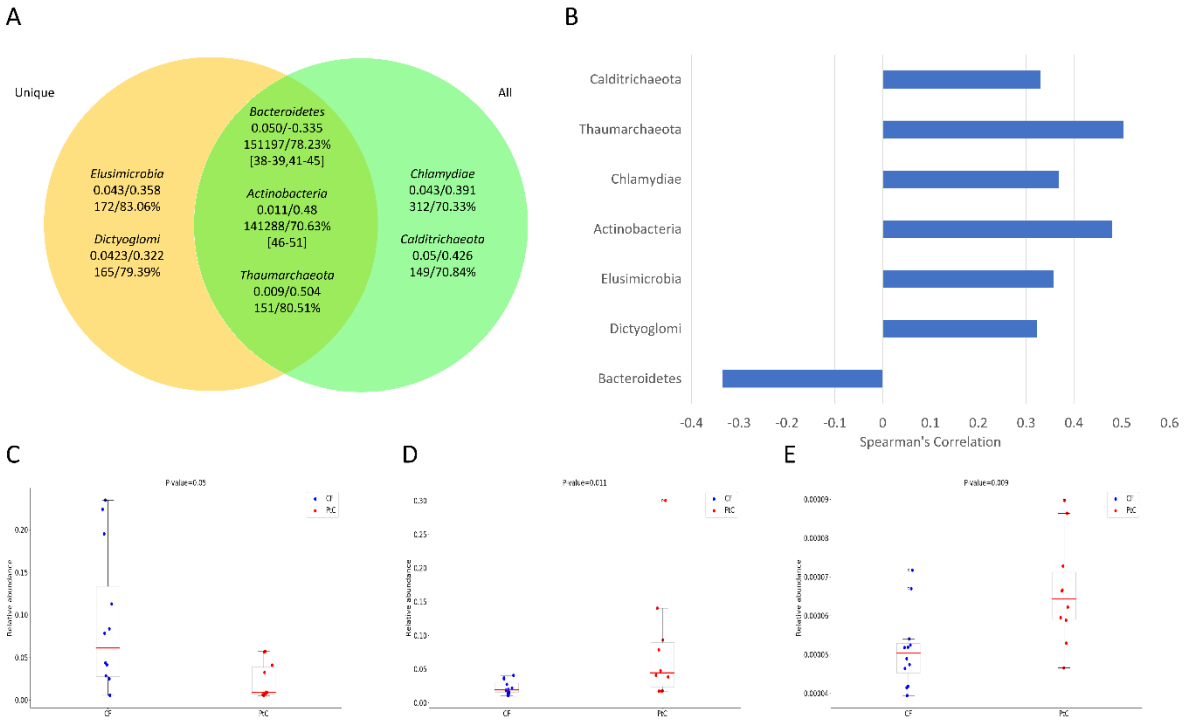

Figure 2 Significant phyla from unique reads only and all mapped reads, respectively.

A. Seven significant phyla identified. The two numbers below a phylum name are the Mann-Whitney p-values, the Spearman's correlations of the read abundances with the colitis severity. The next row provides the average number of reads mapped to each of the 22 samples and the percentage of unique reads among all mapped reads for the phylum. The third row below a phylum name gives the references that may support the colitis-relatedness of this phylum and its lower taxa. B. The Spearman's correlation for seven phyla. C-E: the scatter plot of the relative read abundances in Bacteroidetes, Actinobacteria, and Thaumarchaeota, respectively.

We also studied significant phyla with all mapped reads (i.e., unique reads and multi-reads) (Methods). We identified the following five phyla that are significantly different between samples: Thaumarchaeota (p-value = 0.014), Actinobacteria (p-value = 0.021), Chlamydiae (p-value = 0.043), Calditrichaeota (p-value = 0.050), and Bacteroidetes (p-value = 0.050) (Figure 2A). Bacteroidetes again is not as significant as three of the other four new phyla. The four new phyla also have a higher positive correlation with the colitis severity than Bacteroidetes. The correlations of the read abundances with the severity of colitis for Thaumarchaeota, Actinobacteria, Chlamydiae, Calditrichaeota, and Bacteroidetes are 0.426, 0.484, 0.391, 0.426, and -0.335, respectively. Three phyla (Thaumarchaeota, Actinobacteria, and Bacteroidetes) are identified with unique reads only and with all mapped reads as well, suggesting that at least three phyla may relate to colitis development in patients (Figure 2). The p-value of Thaumarchaeota and Actinobacteria is changed with all mapped reads compared with only unique reads (Figure 2A), indicating the difference of the abundance of multi-reads relative to unique reads between the two types of samples for these two phyla.

Several aspects are different between Bacteroidetes and the six new phyla. First, there are many more reads mapped to Bacteroidetes than to other phyla. In each of the 22 samples, Bacteroidetes on average has 151,197 mapped reads, while the six new phyla except Actinobacteria on average have fewer than 320 mapped reads (Figure 2A). Since the original study applied a marker gene based method to identify significant phyla and it is unlikely that the small number of sequenced reads from the five new phyla come from marker genes, it is not surprising that it missed these five low abundance phyla. In terms of Actinobacteria, which has 141,288 mapped reads on average in each sample, the latest version and the old version of MetaPhlAn indeed identify this phylum

as significant (see the fourth results section). The original study did not report this phylum, maybe because the 16S rRNA read analysis did not show the significance of this phylum. Second, PtC samples have more reads from the six new phyla than CF samples, while it is opposite for Bacteroidetes (Figure 2B). Third, except Dictyoglomi, the abundance of reads from new phyla have more significant correlations with the severity of colitis than that from Bacteroidetes (Figure 2B).

In summary, at least three phyla (Bacteroidetes, Actinobacteria, Thaumarchaeota) are highly likely related to colitis development in patients (Figure 2B-2E). The read abundances of Thaumarchaeota and Actinobacteria are more different between samples compared with Bacteroidetes based on only unique reads and all mapped reads (Figure 2C-2E). Moreover, their abundances correlate with the severity of colitis better than that of Bacteroidetes (Figure 2B). In addition, Elusimicrobia and Chlamydiae may be related to colitis development in patients as well. This is because their properties of read abundances and correlations are similar as the above three phyla, although they are not identified by both all mapped reads and unique reads only. It is worth pointing out that there is at least one significant lower level taxon identified by unique reads from each of these five phyla, as shown in the next section.

2.3.2 Hundreds of lower taxa may relate to colitis development in patients

We further compared read abundances from lower taxa between samples (Methods). If we consider only unique reads, there are 3 classes, 14 orders, 34 families, 101 genera and 244 species with read abundances different between samples (p-value< = 0.05). The original result was saved in supplementary table S1 of (Xin Li, Naser, Khaled, Hu, & Li, 2018). If we consider all mapped reads, 6 classes, 15 orders, 43 families, 116 genera and 334 species have different read abundances between samples (p-value< = 0.05). This part of result was in supplementary S2 of (Xin Li et al.,

2018). In total, there are 7 classes, 20 orders, 52 families, 143 genera and 406 species with read abundances different between samples (Tables 1, Supplementary Table S1 and S2 in original paper) (Xin Li et al., 2018). Note that due to the large number of un-sequenced genomes, when the read abundances of a taxon is significantly different between samples, the read abundances of neither its ancestral taxa nor its offspring taxa may be significantly different between samples.

Table 1 The number of taxa identified based on different criteria

The taxon level	#taxa from unique reads	#taxa from all mapped reads	#taxa from unique or all mapped reads	#correlated taxa from unique reads	#correlated taxa from all mapped reads	#correlated taxa from unique or all mapped reads	#correlated taxa from both unique reads and all mapped reads
phylum	5	5	7	4	5	6	3
class	3	6	7	2	6	6	2
order	14	15	20	12	14	17	9
family	34	43	52	28	40	46	22
genus	101	116	143	95	109	134	70
species	244	334	406	221	309	368	162

The aforementioned five phyla that may relate to colitis development (Thaumarchaeota, Actinobacteria, Elusimicrobia, Bacteroidetes, and Chlamydiae) all have lower taxa that are significantly different between samples based on unique reads. Bacteroidetes has four families, nine genera, and nineteen species with read abundances significantly different between samples. Two of the four families, Rikenellaceae and Barnesiellaceae, which were reported in the original study, are significantly different between samples. Although the abundance of reads from Bacteroidetes itself negatively correlates with the colitis severity, the read abundances from some of its significant lower level taxa positively correlates with the colitis severity. For instance, the species Bacteroides caccae has a p-value of 0.006 and a negative correlation of -0.468, while the

species Blattabacterium sp has a p-value of 0.043 and a positive correlation of 0.438. Actinobacteria has eleven species, two genera, one family, one order and one class with read abundances significantly different between samples. All these lower taxa are all under the class Actinobacteria, which is the class for high G+C Gram-positive bacteria but is not significant itself, implying that certain Gram-positive bacterial species may play an important role in PtC patients. The two phyla, Thaumarchaeota and Elusimicrobia, each has one significant species and at most one significant lower taxon at every lower taxonomical level (Figure 3). For instance, Elusimicrobia has only one class, one order, one family, one genus, and one species with read abundances different between samples. The remaining phylum, Chlamydiae, has one order, one family, two genera and three species with read abundances significantly different between samples (Figure 3).

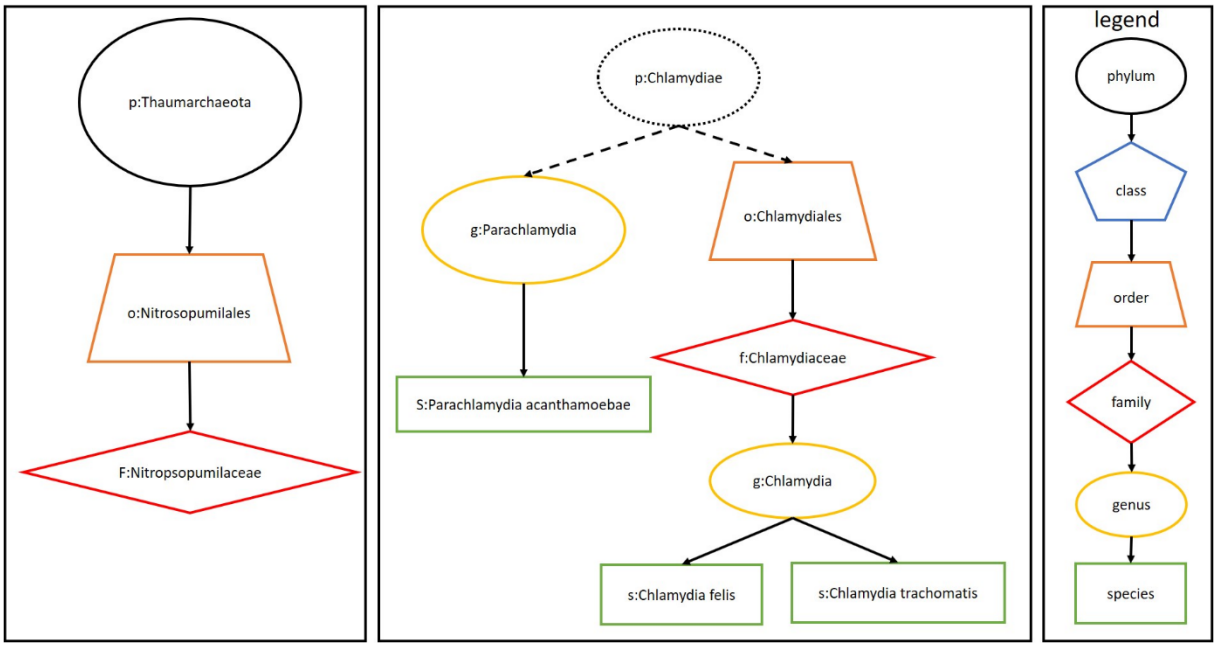

Figure 3 Lower taxa identified from the phyla Thaumarchaeota and Chlamydiae

Only taxa from the last column of Table 1 are shown. Note that no class from these two phyla are identified in the last column of Table 1. The phylum Chlamydiae is presented in a dotted box, as it is not identified in the last column of Table 1

In terms of correlation with the severity of colitis in patients, the read abundances of these significant lower taxa of the above five phyla based on unique reads has higher correlation than that of Bacteroidetes (Supplementary Table S1 in original study) (Xin Li et al., 2018). Bacteroidetes has three families, eight genera and sixteen species with higher correlations than Bacteroidetes itself. Among the three families, Rikenellaceae (p-value 0.021, correlation -0.463) and Barnesiellaceae (p-value 0.025, correlation -0.381) are also identified in the original study, while the family Dysgonamonadaceae (p-value 0.05, correlation-0.342) is missed by the original study. Among all lower taxa, four genera and eight species have positive correlations with the colitis severity, while the remaining eight species, four genera, and three families have negative correlations. The average of these negative and positive correlations is -0.390 and 0.459, respectively. Actinobacteria has ten species, two genera, one family, one order and one class with higher correlations with the severity of colitis, and the average of these correlations is 0.447. Thaumarchaeota has a significant family and a significant order with an average correlation of 0.478 when we consider only unique reads. Elusimicrobia has one significant lower taxon with the correlation of 0.402 at each taxonomical level. Chlamydiae has one order, one family, two genera, and three species with an average of correlation of 0.414.

When we consider all mapped reads, the five phyla discussed above also have lower taxa that are significantly different between samples, and their abundance has higher correlations with the severity of colitis than that of Bacteroidetes (Supplementary S2 in original paper) (Xin Li et al., 2018). Bacteroidetes has three families, nine genera and seventeen families that are significant and have higher correlations. The average negative correlations of lower taxa is -0.381 and the average of positive correlations is 0.433. Actinobacteria has one class, one genus and ten species that are

significant and have an average correlation of 0.456. The same as the lower taxa from unique reads, ten species are from the same class of Actinobacteria, which is the class for gammapositive bacteria and is not significant itself. Another class, Coriobacteriia, is opposite to the Actinobacteria class, in that the class itself is significantly different between the two groups and its abundance correlates better with the severity of colitis while this class has no significant lower taxon. Elusimicrobia has exactly one lower taxon at each level with the abundance significantly different between the two groups and correlating better with the severity of colitis. The average correlation of significant lower taxa is 0.457. Thaumarchaeota has one class, two orders, two families, one genera and one species that are significant and their abundance correlate better with the severity of colitis. The average correlation is 0.396. The class under the Thaumarchaeota phylum, Nitrososphaeria, has exact one taxon at each of its lower levels. Chlamydiae has one class, one order, two families, two genera and four species that are significant and have higher correlations with the severity of colitis, and the average correlation is 0.410. Among them, Chlamydia felis is the species identified with the highest correlation of 0.587 in this study.

It is also worth pointing out that although we focus on lower taxa under the five phyla, there are many significant lower taxa not from these five phyla (Supplementary Table S1 and S2 in original study) (Xin Li et al., 2018). For instance, there are at least 8 orders, 21 families, 82 genera and 192 species that do not belong to the five phyla with their abundance significantly different between the two types of samples and correlate better with the severity of colitis than Bacteroidetes.

A large proportion of the significant lower taxa from unique reads and from all mapped reads are the same (Table 1, Supplementary Table S1 and S2 in original paper) (Xin Li et al., 2018). For all significant lower taxa, we have identified 2 (42.86%) classes, 9 (28.57%) orders, 25 (45.00%) families, 74 (51.75%) genera and 172 (42.36%) species from both unique reads only and from all

mapped reads. All five phyla have at least one shared lower taxon from both unique reads only and from all mapped reads, and the lower taxa have higher correlations than their parent taxa. Among all these significant lower taxa, the abundance of 2 (100%) classes, 9 (100%) orders, 22 (88.00%) families, 70 (94.59%) genera, 162 (94.19%) species correlate with the severity of colitis better than that of Bacteroidetes (Table 1, last column). We believe that these taxa are highly likely colitis-related taxa.

2.3.3 Many identified taxa may relate to colitis development based on literature

Microbes are known to play a vital role in the development of colitis (Campieri & Gionchetti, 2001). We thus studied whether the above taxa are colitis related based on literature, since their abundance is significantly different between samples and correlates better with the severity of colitis than Bacteroidetes. Because the number of these taxa is large, we focused on the three most confident phyla (Bacteroidetes, Actinobacteria, Thaumarchaeota) and their lower taxa. We found that at least 2 (66.7%) of the three phyla and 11 (42.31%) of the identified species in the three phyla are likely colitis related.

We found that microbes from at least two of the three phyla are showed to be related to colitis (Heimesaat et al., 2006; Xu & Jiang, 2017; Ye et al., 2008). Bacteria from Bacteroidetes belong to Gram-negative bacteria (Wexler, 2007), which are known risk factors for inflammatory bowel diseases such as colitis (Bloom et al., 2011; Elinav et al., 2011; Garrett et al., 2007; Vigsnæs, Brynskov, Steenholdt, Wilcks, & Licht, 2012). The phylum Actinobacteria belongs to Gram-positive bacteria. Gram-positive commensal bacteria induce colitis by recruiting colitogenic monocytes and macrophages (Nakanishi, Sato, & Ohteki, 2015). Actinobacteria was found increasingly in abundance in colitis groups compared with control non-colitis groups in different

experiments as well (Frank et al., 2007; Lepage et al., 2011; Nagy-Szakal et al., 2013; Rooks et al., 2014).

There are also four and seven lower taxa that may relate to colitis in the phyla Bacteroidetes and Actinobacteria, respectively. In term of Bacteroidetes, Ye et al. analyzed faecal samples collected from patients with colitis and found that the abundance of Barnesiella viscericola correlates with the disease activity in IL-10-/- mice (Ye et al., 2008). Barnesiella viscericola are found by unique reads with the Mann-Whitney p-value of 0.025 and its abundance has a correlation coefficient of -0.381 with the severity of colitis. Another example is the Bacteroides, whose abundance is significantly different between samples and correlates with the colitis severity better than Bacteroidetes. Bacteroides are found to be accumulated in inflamed ileum at high concentrations (Heimesaat et al., 2006). For all sixteen species in Bacteroidetes with their abundance significantly different between PtC samples and CF samples as well as correlating better with the severity of colitis than Bacteroidetes, three species are from the genus Bacteroides. They are Bacteroides caccae (p-value 0.006, correlation -0.468), Bacteroides salanitronis (p-value 0.025, correlation -0.383) and Bacteroides cellulosilyticus (p-value 0.036, correlation -0.359). As to the phylum Actinobacteria, the analyses of the microbiota in mucosa of patients with ulcerative colitis (UC) show that there are more Actinobacteria and Proteobacteria in patients compared with controls (Lepage et al., 2011). Especially, microbiota of patients with UC have high level of abundance of the genus Rhodococcus and a low abundance of both Bacteroides and Prevotella genera compared with the controls. We found that the abundance of the species Rhodococcus erythropolis under the Rhodococcus genus is significantly different between samples (p-value 0.036, correlation 0.421). Another study also indicates that species in Rhodococcus causes infection in patients (Zinner, 1999). Rooks et al. found that gut microbiomes of colitis patients were most significantly enriched

in Actinobacteria, including Corynebacterium, compared with the controls (Rooks et al., 2014). Four species we identified in this study are from Corynebacterium and have an average correlation of 0.481 with the severity of colitis.

Besides the significant taxa related to colitis from the three most confident phyla, there are other taxa supported by literature (Supplementary Table S1 and S2 in original paper) (Xin Li et al., 2018). These are in total 101 taxa under the phylum Firmicutes (Bartlett, Onderdonk, Cisneros, & Kasper, 1977; Du et al., 2015; Lepage et al., 2011; Nagy-Szakal et al., 2013; Rooks et al., 2014; Vigsnæs et al., 2012; Xu & Jiang, 2017; Zinner, 1999), 3 under the phylum Proteobacteria and 1 under the phylum Fusobacteria (Xu & Jiang, 2017). Among them, the majority of taxa are actually lower level of the Bacillales order, which includes 80 of the lower taxa we identified (the order Bacillales itself, 4 families, 16 genera and 59 species). Rooks et al. demonstrate that Bacillales plays an important role in colitis in gut (Rooks et al., 2014). They also found the genus Staphylococcus are more enriched in colitis patients (Rooks et al., 2014). In our study, we found that the abundance of ten taxa (Staphylococcus itself and nine species) from Staphylococcus are significantly different between samples and correlates well with the colitis severity.

2.3.4 Re-analyses with MetaPhlAn support that Actinobacteria is different between CF samples and PtC samples

The original study generated and analyzed the same shotgun metagenomic dataset with MetaPhlAn (Dubin et al., 2016). From the original study, they concluded that the read abundances of Bacteroidetes and its three families Bacteroidaceae, Rikenellaceae and Barnesiellaceae are significantly different between samples, and correlate well with the severity of colitis. Since we cannot find the list of all taxa this study identified, especially their analyses results from the shotgun metagenomic reads, which was only partially shown in their S2 Fig (Dubin et al., 2016),

we followed their procedure and applied the 1.7.7 version and the 2.1.0 version of MetaPhlAn to the same shotgun metagenomic data (Figure 1). The only difference we made is that we further trimmed reads with Cutadapt to cut common adapters and primers after following their read trim procedure (Martin, 2011).This is because after their suggested read trim procedure from the original study, there are still certain samples with an extremely large ratio of the observed occurrence to the expected occurrence of several k-mers at the beginning or end of reads. The results from two different read trim procedures are actually quite similar, because the number of the affected reads is relative small compared with the number of total reads within samples.

Although we redid the analyses with almost the same procedures by the same tool, our result from both versions of MetaPhlAn is quite different from what was reported in the original study (Table 2). With the old version, MetaPhlAn identified one phylum (Actinobacteria),one class (Actinobacteria), and three species (Alistipes shahii, Clostridium asparagiforme, Bacteroides caccae) with read abundances significantly different between samples (p-value<0.05).The Bacteroidetes phylum itself is not significantly different between samples, although two of the three identified species are from this phylum. The only significant phylum identified is Actinobacteria, together with one of its classes. With the latest version, MetaPhlAn identified one phylum (Actinobacteria), one class (Actinobacteria), one family (Rikenellaceae), one genus(Alistipes), and six species (Alistipes shahii, Alistipes finegoldii, Alistipes onderdonkii, Bacteroides caccae, Eubacterium siraeum, Eubacterium sp. 3_1_31) with read abundances significantly different between samples (p-value<0.05) (Table 2). Similarly, the Bacteroidetes phylum itself is not significantly different between samples, although four of the six identified species together with one identified genus and one identified family are from this phylum. The only significant phylum identified is Actinobacteria, together with one of its classes. One species

identified by the old version is not discovered by the latest version, indicating that multi-reads may affect the downstream analyses and unique reads with current annotation may become multi-reads in the future. We also tried the old version without changing the read trimming procedure in the original research, we still only identified Actinobacteria as the only significant phylum between samples (Table 3). We also studied the correlation of the read abundances of these identified taxa with the colitis severity in patients. All identified taxa have better correlation than Bacteroidetes or almost all of their mapped reads are from one type of samples and thus cannot calculate the correlation (Table 2).

We compared the results from our analyses in the previous sections with those from MetaPhlAn. Many more taxa are identified by mapping reads to available sequenced genomes than by MetaPhlAn (Tables 1 and 2). The reason may be because MetaPhlAn mapped reads to marker genes, which cannot work well when the number of reads from a taxon is limited. Therefore, the two analyses from MetaPhlAn can only identify certain taxa from the two most abundant phyla. In addition, many taxa identified by MetaPhylAn and by the original study are also discovered in our study, supporting the colitis-relatedness of these taxa. A few taxa discovered by MetaPhlAn and by the original study are not found in our study, suggesting that these taxa may be unreliable.

Table 2 Comparison of results from our analyses and from two MetaPhlAn based analyses

	taxa reported by the original study	taxa from MetaPhlAn version 2.7.0	taxa from MetaPhlAn version 1.7.7	taxa from our pipeline that are reported by the original study or identified by MetaPhlAn
Phylum	1(Bacteroidetes)	1(Actinobacteria)	1(Actinobacteria)	2(Bacteroidetes, Actinobacteria)
Class	0	1(Actinobacteria)	1(Actinobacteria)	0
Order	0	0	0	0
Family	3(Bacteroidaceae, Rikenellaceae, Barnesiellaceae)	1(Rikenellaceae)	0	2(Rikenellaceae, Barnesiellaceae)
Genus	0	1(Alistipes)	0	1(Alistipes)

The numbers in the table are the number of significant taxa identified by different pipelines. The names of these taxa are provided in the parentheses.

Table 3 Taxa identified by two versions of MetaPhlAn

Type	name	Corresponding phylum	taxID	score	pvalue	# of CF	# PtC	soi_value	soi_pvalue
class	Actinobacteria	Actinobacteria	1760	94	0.024916	12	10	0.410909	0.05746963
phylum	Actinobacteria	Actinobacteria	201174	94	0.024916	12	10	0.4797036	0.023869957
species	Alistipes_finegoldii	Bacteroidetes	214856	35	0.027812	5	0	-0.4629698	0.030022629
species	Alistipes_onderdonkii	Bacteroidetes	328813	31	0.035001	7	2	None	None
species	Alistipes_shahii	Bacteroidetes	328814	27.5	0.014375	7	1	None	None
species	Bacteroides_caccae	Bacteroidetes	47678	30	0.041087	8	4	-0.467928	0.028081632
species	Eubacterium_siraeum	Firmicutes	39492	32	0.041914	7	2	None	None
species	Eubacterium_sp_3_1_31	Firmicutes	457402	87.5	0.023312	1	5	None	None

2.4 Discussion

By mapping metagenomic reads to all available microbial genomes, we identified at least 3phyla, 2 classes, 9 orders, 22 families, 70 genera and 162 species that are potentially colitis related(last column of Tables 1, last column of Supplementary Table S1 and S2 in original paper) (Xin Li et al., 2018). This is because the abundance of each of these identified taxa is significantly different between CF and PtC samples, and correlates with the colitis severity in patients better than the abundance of Bacteroidetes. Moreover, these taxa are identified by both unique reads and all mapped reads. In addition, 2 phyla, 1 order, 4 families,18 genera and 71 species are colitis-related based on literature search (Supplementary Table S1 and S2 in original paper) (Xin Li et al., 2018). Compared with the previously identified colitis-related taxa from the same data, we identified much more taxa supported by literature.

We require that the read abundances of potential colitis-related taxa is significantly different between CF and PtC samples, and correlates well with the colitis severity, for both unique reads only and for all mapped reads together. We have lower confidence on the colitis-relatedness of certain taxa such as the Chlamydiae phylum, although its abundance of all mapped reads instead of only unique reads is significantly different between samples, and correlates well with the colitis severity. This is because of our assumption that reads are randomly chosen to be sequenced from a genome and there should be more unique regions for a given microbial genome than shared regions with other genomes. Under this assumption, a significant taxons hould have unique read abundances significantly different between CF and PtC samples.

We show that multi-reads affect the analysis results. The inferred taxa based on unique reads only are not always consistent with and sometimes quite different from the inferred ones based on all

mapped reads. This implies the necessity to develop better methods to accurately assigned multi-reads to the [a]bona fide[o] genomes, which cannot be done satisfactorily at present. Moreover, this also calls for cautious consideration when we remove duplicated reads before mapping. Different from read mapping in individual species, where duplicated reads only affect a small portion of repetitive regions, duplicated reads in metagenomics likely affect the analysis of the present species and their abundance, as duplicated reads can be mapped to multiple species as well.

Although we do not have high confidence on the colitis-relatedness of certain taxa because they are insignificant based on unique reads, they can be still biologically significant and related to colitis development. For instance, the Chlamydiae phylum is not considered colitis related in our study. However, its lower taxa at the level of order, family, genus, and species are all significant based on unique reads. The abundance of these significant lower taxa has an average correlation with the severity of colitis around 0.41. One of its lower taxa at the species level, Chlamydia felis, has a correlation of 0.415. Although zoonotic infection of humans with Chlamydia felis is not reported, Chlamydia felis is a bacteria found in cats and is primarily for the inflammation of feline conjunctiva, rhinitis and respiratory problems (Azuma et al., 2006; Everett, Bush, & Andersen, 1999).

Table 4 The number of mapped and unmapped reads in the 22 samples

cellName	# of raw reads	# of reads after common cut	# of reads after further cut	# of unique mapped reads	# of all mapped reads	# of unmapped reads	percentage of unmapped reads
Feces_pt_PtC_34	2023057	1977088	1967727	509143	666842	1356215	67.04%
Feces_pt_PtC_26	1727537	1638429	1630798	768247	887518	840019	48.63%
Feces_pt_PtC_25	2255027	2250710	2245517	490404	671259	1583768	70.23%
Feces_pt_PtC_30	1247190	1242742	1237690	497615	592763	654427	52.47%
Feces_pt_PtC_31	1805564	1722569	1716466	389883	495201	1310363	72.57%
Feces_pt_PtC_32	3338636	3305876	3265887	933375	1350585	1988051	59.55%
Feces_pt_PtC_33	1800362	1754685	1748414	416178	519046	1281316	71.17%
Feces_pt_PtC_29	1829492	1788929	1781507	548619	663644	1165848	63.73%
Feces_pt_PtC_28	1563526	1518722	1510683	390619	493493	1070033	68.44%
Feces_pt_CF_21	2028384	1993731	1986029	820419	1061438	966946	47.67%
Feces_pt_CF_17	2155910	2138789	2124925	422931	558776	1597134	74.08%
Feces_pt_CF_15	1980883	1936329	1929169	820823	1014835	966048	48.77%
Feces_pt_CF_14	1925574	1880572	1873141	688959	823779	1101795	57.22%
Feces_pt_CF_10	3502444	3475330	3448720	1180936	1793658	1708786	48.79%
Feces_pt_CF_22	1324814	1282050	1276577	405975	513199	811615	61.26%
Feces_pt_CF_23	2817535	2811737	2804310	997574	1144696	1672839	59.37%
Feces_pt_CF_20	1930172	1925674	1920959	484927	585052	1345120	69.69%
Feces_pt_CF_18	2532048	2526046	2517608	614702	796244	1735804	68.55%
Feces_pt_PtC_27	3110350	3085632	3063954	756410	963197	2147153	69.03%
Feces_pt_CF_7	1778795	1739830	1733110	516251	624455	1154340	64.89%
Feces_pt_CF_6	3998572	3971403	3945823	1166643	1495820	2502752	62.59%
Feces_pt_CF_2	1313013	1278494	1272123	569519	736458	576555	43.91%
Average	2181313	2147517	2136415	654098	838725	1342588	61.35%

We compared our results based on sequenced genomes with those from MetaPhlAn. We identified many more colitis-related taxa based on sequenced genomes (Table 1). The majority of these missed taxa by MetaPhlAn analyses are low abundant. They are missed by MetaPhlAn, likely because there are many fewer reads that can be mapped to marker genes by MetaPhlAn and thus these low-abundant taxa are not different between CF samples and PtC samples. In addition, since the original study was submitted in November 2015, there are 74 (39.15%) and159 (38.50%) species sequenced in Bacteroidetes and Actinobacteria, respectively. We found that the read abundances of 5 of the 74 species and that of 4 of the 159 species are significantly different between CF samples and PtC samples (Supplementary Table S1 and S2 in original study) (Xin Li et al., 2018), which cannot be identified by MetaPhlAn, as the latest version of MetaPhlAn does not include these species. With more sequenced genomes in the future, with our pipeline or with MetaPhlAn, we may identify even more colitis-related species, as there are on average only about 38.65% reads that can be mapped to the sequenced genomes currently (Table 4). It is worth pointing out that, unexpectedly, different from what the original study reported, the application of two versions of MetaPhlAn shows that Actinobacteria instead of Bacteroidetes has significantly different abundance between samples (Table 2), suggesting that 16S rRNA read analyses resulted in a different set of taxa from the analyses based on MetaPhlAn. Such an unexpected difference also implies the limitation of 16S rRNA profiling based approaches.

Our study shed new light on metagenomic studies. It shows the necessity to consider every region in sequenced genomes instead of considering marker genes only. It also suggests caution when working with duplicated reads and multi-reads during the analyses. Moreover, it is mandatory to take into account how newly sequenced genomes affect the results if methods based on sequenced

genomes are used. We hope that in the near future, new and better tools to consider multi-read mapping and novel methods without relying on sequenced genome scan be developed so that the issues here can all be addressed or at least minimized.

CHAPTER THREE: AN APPROACH FOR BACTERIAL STRAIN RECONSTRUCTION BASED ON DE BRUIJN GRAPH

Previously published as Li, X., Saadat, S., Hu, H., & Li, X. (2019). BHap: a novel approach for bacterial haplotype reconstruction. Bioinformatics, 35(22), 4624-4631.

3.1 Introduction

It is important to reconstruct strains from bacterial clonal populations. Strains are variant copies of a genome in a population that are created gradually with accumulated mutations in DNA (Lang, Botstein, & Desai, 2011). Reconstructing strains in a bacterial population reveals the population structure and its evolutionary features (Pulido-Tamayo et al., 2015). In addition, reconstructing bacterial strains is required to choose the right treatments for diseases caused by specific strains in a population, which may vary in only a few base pairs (bps) compared with other strains in the population (Schirmer, 2014).

The next generation sequencing (NGS) technologies provide a unique opportunity to reconstruct strains in bacterial clonal populations. NGS technologies can sequence DNA from a bacterial population (Barrick & Lenski, 2009). The sequenced short reads are a mixture of DNA segments from different strains in the population. Researchers can then regroup reads for individual strains, reconstruct the strains and discover the diversity in the population from these reads.

Several approaches have been developed for viral strain reconstruction. ShoRAH performs a local analysis to estimate the strain diversity at the local level and then applies a global analysis using the path cover algorithm to reconstruct genome-wide strains (Zagordi et al., 2011). QColors constructs the read conflict graph and models the population reconstruction as a vertex coloring

problem (Huang, Kantor, DeLong, Schreier, & Istrail, 2011). ViSpA creates a weighted overlap graph for reads and iteratively finds maximum-weight paths and considers them as viral strains (Astrovskaya et al., 2011). QuRe partitions the reference genome with mapped reads into sliding windows and scores partitions, and then constructs an overlap graph, and finally finds the path of genomes by a heuristic algorithm (Prosperi & Salemi, 2012).

Although the aforementioned methods are suitable for viral populations, they have difficulty in distinguishing bacterial strains, which are more similar to each other due to much lower mutation rates compared with those in viral populations. In viral populations, the genomic distance between polymorphic sites is often shorter than the read length. Every read may thus contain polymorphic sites. Moreover, overlapping reads from the same strain likely share common polymorphic sites. Viral strain reconstruction methods typically use such overlapping information to infer strains. However, using this piece of information is not enough for bacterial population reconstruction due to the much lower number of mutations. The distance between polymorphic sites in bacterial genomes is often longer than several thousand bps (Pulido-Tamayo et al., 2015). In other words, many reads contain no polymorphic site. Read overlapping thus cannot facilitate the grouping of reads with adjacent polymorphic sites in strains.

To our knowledge, EVORhA is the first and the only existing strain reconstruction tool capable of identifying strains in bacterial populations (Pulido-Tamayo et al., 2015). It defines windows on aligned short reads and infers template strains per window to construct strains locally. It then extends windows by concatenating template strains based on their shared polymorphic sites. EVORhA reconstructs the final genome-wide strains using the relative coverage of the extended strains. Such a local-extension based strategy may be affected by ‘errors’ at the local levels and generates many false positive strains.

In this study, we propose a strain reconstruction method for bacterial populations called BHap (Abbreviation for Bacterial Haplotype Reconstruction). Different from previous studies (Prosperi & Salemi, 2012; Pulido-Tamayo et al., 2015; Zagordi et al., 2011), which often start from locally constructed strain segments and then extend these segments to obtain final strains, BHap always focuses on all polymorphic sites in a strain instead of local genomic regions, by an Expectation-Maximization (EM) algorithm and a fuzzy flow approach. Such a global-based approach, with a guidance of the estimated 'global' picture of the strain coverage, may be more robust to 'errors' and biases in local genomic regions. Tested on simulated and experimental datasets, BHap is capable of reliably reconstructing strains with an average F1 score of 0.87, an average precision of 0.86 and an average recall of 0.88. Compared with existing approaches, BHap constructs more accurate strains and generates fewer false positive strains. The BHap tool is available on http://www.cs.ucf.edu/~xiaoman/BHap/.

3.2 Materials and Methods

3.2.1 Simulated datasets

To investigate the performance of BHap, we simulated 339 datasets with different configurations, such as different coverage, read lengths, mutation rates, strain proportions and sequencing error rates (Table 5). Coverage refers to the sequencing depth of a dataset. It is defined as the ratio of the sum of the length of all reads in a dataset to the length of the corresponding reference genome. The main reason to test BHap on simulated instead of experimental datasets is that polymorphic sites are known in simulated datasets while unavailable in experimental ones, which are essential for an accurate evaluation of the methods.

To simulate data, we randomly selected the genomes of three bacterial species, Bartonella clarridgeiae, Enterococcus casseliflavus and Methanobrevibacter smithii (GenBank NC_014932,

NC_020995 and NC_009515, respectively), as reference genomes. For each of the three reference genomes, we generated a default population composed of two strains with the default parameters (Table 5). Since bacterial populations often contain mutations several thousand bps apart from each other, the default mutation rate was set to be 0.01% (Pulido-Tamayo et al., 2015). Here the mutational rate is the percentage of the variations in a strain when it is compared with its reference genome. For every strain in a population, we simulated short paired-end Illumina reads using the dwgsim tool (https://github.com/nh13/DWGSIM) (Table 5). All simulated reads for all strains in the same population were mixed together as a simulated dataset to infer the original strains in this population.

Table 5 Detail simulation dataset information

Group ID	description	number of strains	proportions	mutation rate	read length	error rate	coverage	# of datasets
1	Default parameter for three species	2	30/70	0.01%	100	0.001	100x	15
2	Reads length from 60bp to 150bp except 100bp	2	30/70	0.01%	60-150 except 100bp	0.001	100x	27
3	Sequence error from 0.2% to 1.5%	2	30/70	0.01%	100	0.2%-1.5%	100x	42
4	Proportions 10/90, 20/80, 40/60, 50/50	2	10/90;20/80;40/60;50/50	0.01%	100	0.001	100x	12
5	10/90 for coverage 100x, 200x, 300x, 400x and 500x	2	10/90	0.01%	100	0.001	100x,200x,300x,400x,500x	15
6	Mutation rate is from 0.02% to 0.05%	2	30/70	0.02%-0.05%	100	0.001	100x	12
7	Coverage 50x, 100x, 150x and 200x	2-4	30/70;10/30/60;10/20/30/40	0.01%	100	0.001	50x,100x,150x,200x	36
8	Mutation rate are 0.02% and 0.05% on three proportion	2-4	30/70;10/30/60;10/20/30/40	0.02%, 0.05%	100	0.001	500x	18
9	Mutation rate is 0.07%, 0.1% and 0.15%	2	30/70	0.07%,0.1%,0.15%	100	0.001	50x,100x,150x,200x,500x	45
10	3 and 4 strains with evolutionary relationship	3,4	10/30/60;5/25/70;5/15/25/55;5/10/35/50	0.01%	100	0.001	200x,300x,400x,500x	72
11	Mutation rate from 0.01% to 0.05% on two proportions with three types of evolutionary relationship	3,4	2/25/70; 5/10/35/50	0.01%-0.05%	100	0.001	300x	45

We used dwgsim to simulate reads for each strain with the following commands:
Group1: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C 30.0 -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName
Group2: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C 30.0 -1 read_len -2 read_len -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName
Group3: dwgsim -e sequence_error/2 -E sequence_error/2 -d 400 -s 50 -N -1 -C 30.0 -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName

Group4: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C coverage -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName
Group5: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C coverage -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName
Group6: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C coverage -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName
Group7: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C 30.0 -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome_with_corresponding_mutation resName
Group8: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C coverage -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName
Group9: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C coverage -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome_with_corresponding_mutation resName
Group10: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C coverage -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome resName
Group11: dwgsim -e 0.00050 -E 0.00050 -d 400 -s 50 -N -1 -C 300.0 -1 100 -2 100 -r 0 -z 1534513916 -n 0 -X 0 -R 0 -c 0 -S 0 -y 0 genome_with_corresponding_mutation resName
The resName above is the name of the file to store the generated reads. For each strain proportion in groups 10-11, when there are four strains, we considered two different evolution trajectories for each given strain proportion.
For training datasets to determine the k-mer length automatically, we totally simulate 150+45+75=270 datasets:
150 datasets: read length from 60bp to 150bp on coverage 50x, 100x, 150x, 200x and 500x.
45 datasets: three proportions (30/70,10/30/60 and 10/20/30/40) on coverage 50x,100x,150x,200x and 500x.
75 datasets: five proportions (10/90,20/80,30/70,40/60 and 50/50) on coverage 50x,100x,150x,200x and 500x.

To study how different parameters affect the performance of BHap, we simulated eleven groups of datasets (Table 5). The first group consisted of the above three default populations together with twelve populations generated similarly with the default parameters. In each of the ten remaining groups, the value of one or more parameters was changed. The second group was to study the effect of the read length on the strain reconstruction, in which we changed the read length from 60 to 150 bps except the default 100 bps for each of the above three reference genomes. The third group contained 42 datasets, where sequencing error rates varied from 0.2 to 1.5% for each of the three default populations. There were twelve datasets in the fourth group, with four individual strain proportions for each of the three default populations. A strain proportion tells the percentage of reads from every strain. For instance, the strain proportion 10/30/60 tells that 10, 30 and 60% of reads are from three strains, respectively. This fourth group of datasets was used to assess the performance of BHap in reconstructing individual strains with different strain proportions. To study the BHap performance with a 10/90 proportion on higher coverage, we generated additional fifteen datasets with different coverage as the fifth group. Since the mutation rate may affect the strain reconstruction, we simulated twelve datasets in the sixth group with the mutation rates ranging from 0.02 to 0.05%. We generated three additional groups with 99 datasets to compare BHap with the only existing method for bacterial strain reconstruction, EVORhA (Pulido-Tamayo et al., 2015). In the seventh group, for each of the three bacterial genomes, with the coverage of 50×, 100×, 150× and 200× and with each of the following three strain proportions: 30/70, 10/30/60 and 10/20/30/40, we generated twelve datasets. Since the best performance of EVORhA happened at higher coverage, we also generated the eighth group with additional six datasets for each bacterial genome, with 500× coverage, two different mutation rates and three strain proportions. Since the mutation rate in the EVORhA study was higher, we generated the ninth group with the

mutation rate as 0.07, 0.1 and 0.15% and the coverage as 50×, 100×, 150×, 200× and 500× for each of the three genomes, respectively. Since strains in a population evolved from the same reference genome through different trajectories, we simulated two additional groups with 117 datasets so that an evolutionary relationship was enforced in three or four strains in each dataset (the tenth and the eleventh). Note that the enforced evolution relation on a population with only two strains was not meaningful, since they were equal to those we already studied in the first nine groups. We thus studied three evolution trajectories for populations with three and four strains. In brief, we simulated populations with three or four strains, different strain proportions and different mutation rates. For populations with three strains, two strains were set to share a given fraction of polymorphic sites, and the remaining one share no polymorphic site with first two (Type 0 evolution trajectory). For populations with four strains, we considered two different evolution scenarios: Type 1: Two strains share a fraction of their polymorphic sites, the third share fewer polymorphic sites with the first two and the fourth share no polymorphic site with the first three; and Type 2: The first two strains share a fraction of polymorphic sites and the remaining two share a fraction of polymorphic sites, while the two pairs share no polymorphic site (Table 5).

3.2.2 Experimental datasets

We tested BHap on two experimental datasets: the mixed infection dataset of Clostridium difficile and the evolved population dataset of Escherichia coli strain SX4. The mixed infection dataset was generated with the Illumina technology at 150× coverage (Eyre et al., 2013). There were 54 mixed samples, each constructed from two of the 36 unmixed samples (https://www.ebi.ac.uk/ena/data/view/PRJEB1729). The two unmixed samples used to construct a mixed sample and their proportions were provided in Supplementary S2 of original paper (Xin Li, Saadat, Hu, & Li, 2019). For the evolved population dataset, 100 bps long paired-end Illumina

reads at a coverage of ~200× were available at three time points (http://www.ncbi.nlm.nih.gov/bioproject/262000). At every time point, a population as well as a corresponding clone were sequenced. Here the number and the strain proportions were unknown, while the strain(s) in the clone was likely in the population at the corresponding time points.

For the mixed infection dataset, we ran BHap and EVORhA on each mixed sample and each of its two corresponding unmixed samples to predict strains. We then calculated the similarity of every pair of predicted strains, with one strain from a mixed sample and the other strain from its unmixed samples. The similarity of a pair of strains u and v was calculated in exactly the same way as that in EVORhA, which was calculated as (1), where P_u and P_v are the set of polymorphic sites in u and v, respectively. This similarity was called reliability in the EVORhA study. In this way, we identified one pair of most similar strains for a mixed sample and each of its unmixed samples. Finally, we averaged the reliability of these two pairs of strains for a mixed sample and its two unmixed samples to measure the performance of BHap and EVORhA. Similarly, for the evolved population dataset, at each time point, we predicted strains in the population sample and its corresponding clone sample using BHap and EVORhA. We then identified the most similar pair of predicted strains with one from the population sample and the other from its corresponding clone sample. Finally, we output the similarity of the most similar pair of strains to measure the reliability of BHap and EVORhA at each of the three time points.

$$\boldsymbol{reliability} = \frac{|\boldsymbol{P_u} \cap \boldsymbol{P_v}|}{|\boldsymbol{P_u}| + |\boldsymbol{P_v}|} \qquad (1)$$

Alternatively, we defined polymorphic sites in the clone or unmixed samples by SAMtools, which is commonly used to infer polymorphic sites from NGS reads (H. Li, 2011; H. Li et al., 2009). We then applied BHap and EVORhA to the corresponding population (mixed) samples to predict strains. Finally, we compared the polymorphic sites in the predicted strains with those inferred

from SAMtools to calculate the reliability of BHap and EVORhA. This was to make sure that BHap or EVORhA identified those polymorphic sites in populations that were also discovered independently in the corresponding clones or unmixed samples.

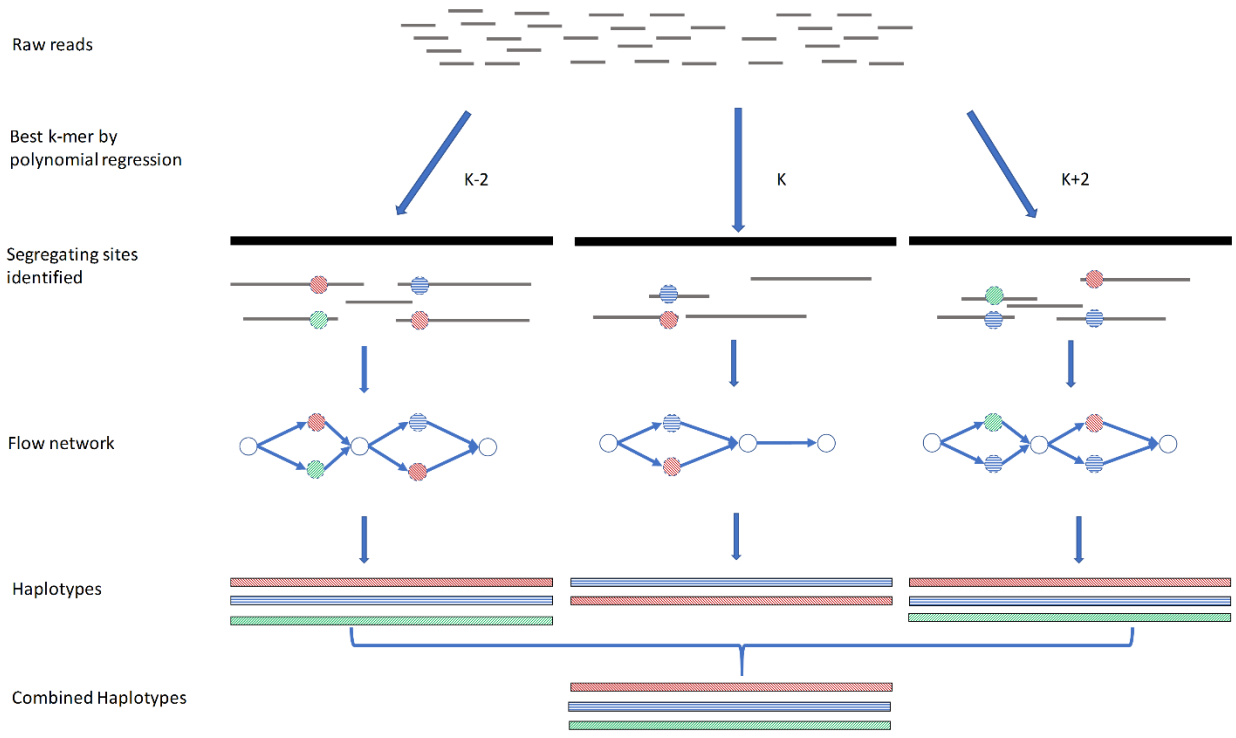

Figure 4 Flowchart of the BHap algorithm

Polymorphic nodes from different strains and the strains themselves are drawn with different patterns

3.2.3 BHap, a novel approach for strain reconstruction in bacterial populations

BHap is composed of the following four major steps (Figure. 4): it determines a proper k-mer length for constructing a de Bruijn graph; it then creates a flow network from the de Bruijn graph and identifies sequencing errors and polymorphic sites; next, it decomposes the flow network to

infer feasible flows with an EM algorithm. These flows are considered as potential strains; finally, it repeats the above three steps with different k values and combines the results to infer the final strains. See Sections 2.3.1–2.3.5 for details.

3.2.3.1 Choosing a proper k-mer length

The k-mer length affects the construction of the de Bruijn graph and the inference from this graph (Zerbino & Birney, 2008). We observed that for the same k-mer length, when the read length, coverage or the reference genome size is different, our earlier BHap versions had different specificity and sensitivity and thus different F1 scores, in terms of correctly grouping polymorphic sites for individual strains. To automatically choose a proper k-mer length, we trained the following polynomial regression model with 270 simulated datasets: $y_i = a_0 + \sum_{j=1}^{3} a_j x_{ij} + \sum_{j=1}^{3} \sum_{l=1}^{3} b_{jl} x_{ij} x_{il}$, where y_i is the k-mer length that resulted in the best F1 score for the i-th dataset; $x_{ij}, j = 1, 2, 3$ is the average read length, the coverage, the genome size in the i-th dataset, respectively; and the remaining variables are unknown parameters to be estimated from the regression. These 270 simulated datasets were generated similarly as the simulated datasets in Table 5 with the dwgsim tool, the three reference genomes and the following parameters: seven different strain proportions (10/90, 20/80, 30/70, 40/60, 50/50, 10/30/60, 10/20/30/40), ten different read lengths from 60 to 150 bps, and five different coverage (50×, 100×, 150×, 200×, 500×). Given a new dataset, $x_{ij}, j = 1, 2, 3$ are known and the best k-mer length is obtained from the above model with the estimated parameters by the BHap tool.

3.2.3.2 Construction of the flow network

To identify polymorphic sites, BHap applies Velvet (Zerbino & Birney, 2008) to construct a de Bruijn graph and then converts this graph into a flow network. Velvet is a popular tool for

assembling NGS reads based on de Buijn graphs. The de Bruijn graph is a time and memory efficient data structure commonly used to represent short reads for sequence assembly.

For a given k-mer length, each node in the de Bruijn graph represents a k-mer in input reads and each directed edge represents a (k + 1)-mer in input reads. In other words, each edge connects two nodes representing the two k-mers contained in the corresponding (k + 1)-mer for this edge. Edges are weighted with the corresponding number of reads containing the corresponding (k + 1)-mer.

With the de Bruijn graph, BHap applied Velvet to generate uncorrected contigs (Zerbino & Birney, 2008). These contigs are constructed without sequencing error correction and some contigs may thus have low coverage. This is different from normal assembly, in which sequencing error correction is carried out before assembly and corrected contigs are produced as the final product. BHap considers only the uncorrected contigs produced by Velvet, since the goal is to identify polymorphic sites.

BHap then constructs a flow network with the uncorrected contigs. BHap constructs one node in the network for every contig with coverage larger than a specified threshold (default 3). The coverage of contigs is calculated by Velvet, representing its estimated sequencing depth. The contigs with coverage smaller than the threshold likely contain sequencing errors and the remaining contigs likely contain all polymorphic sites. For each node, BHap maps the corresponding contig sequence to the reference genome by the BLAT tool (Kent, 2002). BLAT is used since the contigs are relative long and their number is much fewer compared with the input reads. If two contigs are mapped to overlapping regions, BHap connects the two corresponding nodes with a directed edge, in the same order as their occurrence in the reference genome. To reduce the storage cost, BHap merges consecutive nodes in a path that are not shared by any other

path into one node. The edge weights are modified with the coverage of the corresponding sequences. In this way, BHap constructs a flow network, with the coverage of nodes as the flow capacities. In this network, nodes immediately following the branching points are likely polymorphic sites, which are called polymorphic nodes in the following (Figure 4).

3.2.3.3 An EM algorithm for finding the capacities of an initial flow set

The capacity of nodes from each strain can be approximated as a Poisson distribution (Ying Wang et al., 2015). Therefore, the coverage of the nodes in the population is a mixture of different Poisson distributions. Since polymorphic nodes distinguish different strains, BHap applies an EM algorithm on the capacity of polymorphic nodes to find an initial set of Poisson distributions and flows.

In brief, assume that X = $\{x_1, x_2, \ldots, x_n\}$ is the list of the polymorphic node capacities that follows a mixture of m Poisson distributions with unknown parameters M = $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, where mis unknown. In the Expectation step, BHap calculates the mean of the missing variables Z_{ij} using (2) and (3), where pr(λ_r, x_i) for r from 1 to m, y_i is an indicator function, y_i = j means that xi is from the j-th Poisson distribution, and αj is the unknown probability that the capacity of a random node is from the j-th distribution. The Maximization step is estimating the parameters using α_j and λ_j, Equation (4). The unknown parameter m is inferred similarly as in a previous study by starting from a large m and decreasing m by one at a time until the obtained m groups of parameters had no two groups with highly similar parameters (Xiaoman Li & Waterman, 2003). The polymorphic nodes are correspondingly grouped based on their probabilities of belonging to the m groups calculated with the final inferred $\{\alpha_j, \lambda_j, j = 1, \ldots, m\}$

$$\boldsymbol{Z_{ij} = Pr(y_i = j | X, \theta) = \frac{\alpha_i \times p_j(\lambda_j, x_i)}{\sum_{r=1}^{m} \alpha_r \times p_r(\lambda_r, x_i)}} \qquad (2)$$

$$\boldsymbol{p_r(\lambda_r, x_i) = \frac{\lambda_r^{x_i} e^{-\lambda_r}}{x_i!}} \quad (3)$$

$$\boldsymbol{\alpha_j = \frac{1}{n}\sum_{i=1}^{n} Z_{ij}}\ , \qquad \boldsymbol{\lambda_j = \frac{\sum_{i=1}^{n} Z_{ij} x_i}{\sum_{i=1}^{n} Z_{ij}}} \quad (4)$$

3.2.3.4 Fuzzy flow decomposition for finding strains

Now BHap tries to decompose the network into a set of flows F={f_1, f_2,…, f_k}, where k is the number of flows (strains) and f_i is the coverage of the i-th strain for i from 1 to k. Under the assumption that different strains have different coverage, the smallest parameter in M, λ_{min}, is likely the coverage of a strain, while other parameters in M may be the sum of the coverage of different strains. BHap thus intends to first identify the strain with the coverage of λ_{min}.

To obtain this strain, BHap calculates the cost of passing a flow with the capacity of λ_{min} through each polymorphic node. There is no need to calculate the cost of going through other nodes since they are shared by all strains. The cost of passing a polymorphic node is calculated as (5) and (6), where x is the coverage of this polymorphic node. BHap then identifies the path with the lowest cost that covers the reference genome and output the first strain.

$$\text{cost} = \boldsymbol{1 - \max_{\lambda_j \neq \lambda_{min}} p_j(\lambda_j, x - \lambda_{min})} \quad (5)$$

$$\boldsymbol{p_j(\lambda_j, x - \lambda_{min}) = \frac{e^{-\lambda_j} \lambda_j^{x-\lambda_{min}}}{(x-\lambda_{min})!}} \quad (6)$$

BHap subtracts the capacity of the polymorphic nodes by λ_{min}, for nodes in the latest extracted path. BHap then repeats the procedure of applying the EM algorithm on the polymorphic nodes with the remaining capacities, identifying λ_{min} in the updated M, outputting the flow and path with the capacity of the updated λ_{min}. The algorithm stops when the network becomes disconnected or when the cost of the current path exceeds a specific threshold.

3.2.3.5 Combining results of different k-mer values

Strains in a population may have different coverage. Previous studies show that strains with different coverage can be assembled better with different k-mer lengths (Surget-Groba & Montoya-Burgos, 2010). BHap thus uses different k-mer lengths to reconstruct strains and clusters strains from different k values to find the final set of strains. BHap first selects a proper k-mer length by the above polynomial regression. BHap then considers two additional k-mer lengths that are larger or smaller than this k-mer length by 2. Such a combination showed the best F1 scores on the above 270 simulated datasets used to determine k.

With three k-mer lengths, BHap obtains three sets of strains. Since the same strain in different sets should have similar coverage, BHap assigns strains of similar coverage from different sets to the same cluster. BHap considers every strain in the set resulted from the best k-mer length to be a different initial cluster. For each remaining strain set, BHap compares the coverage of each of its strains with the coverage of the existing clusters. The coverage of a cluster is the average coverage of its strains. If for a strain, difference between its coverage and the coverage of the cluster with the most similar coverage is larger than half of the average coverage difference of the existing clusters, the algorithm creates a new cluster for this strain. If one strain has the same coverage difference when compared with two clusters, the algorithm assigns the strain to the cluster that has more shared polymorphisms with this strain. After assigning strains in a strain set, the algorithm updates the coverage of the clusters and continues to work on another set. With the final clusters of strains, BHap finds the consensus strain in each cluster, with its polymorphisms as those shared by the majority of strains in this cluster.

3.2.4 Evaluation of BHap and other tools

We used precision, recall and F1 score to assess the performance of BHap and other tools on simulated data. On experimental data, where the strains were unknown and these measurements could not be calculated, we used the reliability defined by the EVORhA study instead (Pulido-Tamayo et al., 2015).

3.3 Results

3.3.1 BHap has a robust performance with varied parameter values

To evaluate BHap, we compared the BHap predicted strains with known strains on simulated datasets (Material and Methods). We found the corresponding known strain for each predicted strain. We then compared the polymorphic sites in known strains with those in the corresponding predicted strains. In each dataset, to measure the performance of BHap, we calculated the precision, recall and F1 score based on the predicted polymorphic sites compared with the corresponding known ones. BHap had a good and robust performance in almost all cases.

Under the default parameters, BHap had a recall of 0.88, a precision of 0.86 and an F1 score of 0.87 (Figure 5A and Table 6). Such an average performance was based on 15 simulated datasets with two strains in each dataset. In these datasets, the default average read length was 100 bps and the default sequencing error rate was 0.1%, which mimicked the parameters from the Illumina sequencers (Glenn, 2011). The default coverage was 100×, which was realistic in current practice with significantly decreased sequencing cost.

We studied how the performance of BHap varied with different read lengths (Figure 5B and Table 3). The three measurements, especially the F1 score, were close to each other with varied read lengths. The largest F1 score appeared at the read length of 90 bps, and slightly decreased if we

increased or decreased the read length. We thus concluded that the read length has a limited effect on the performance of BHap.

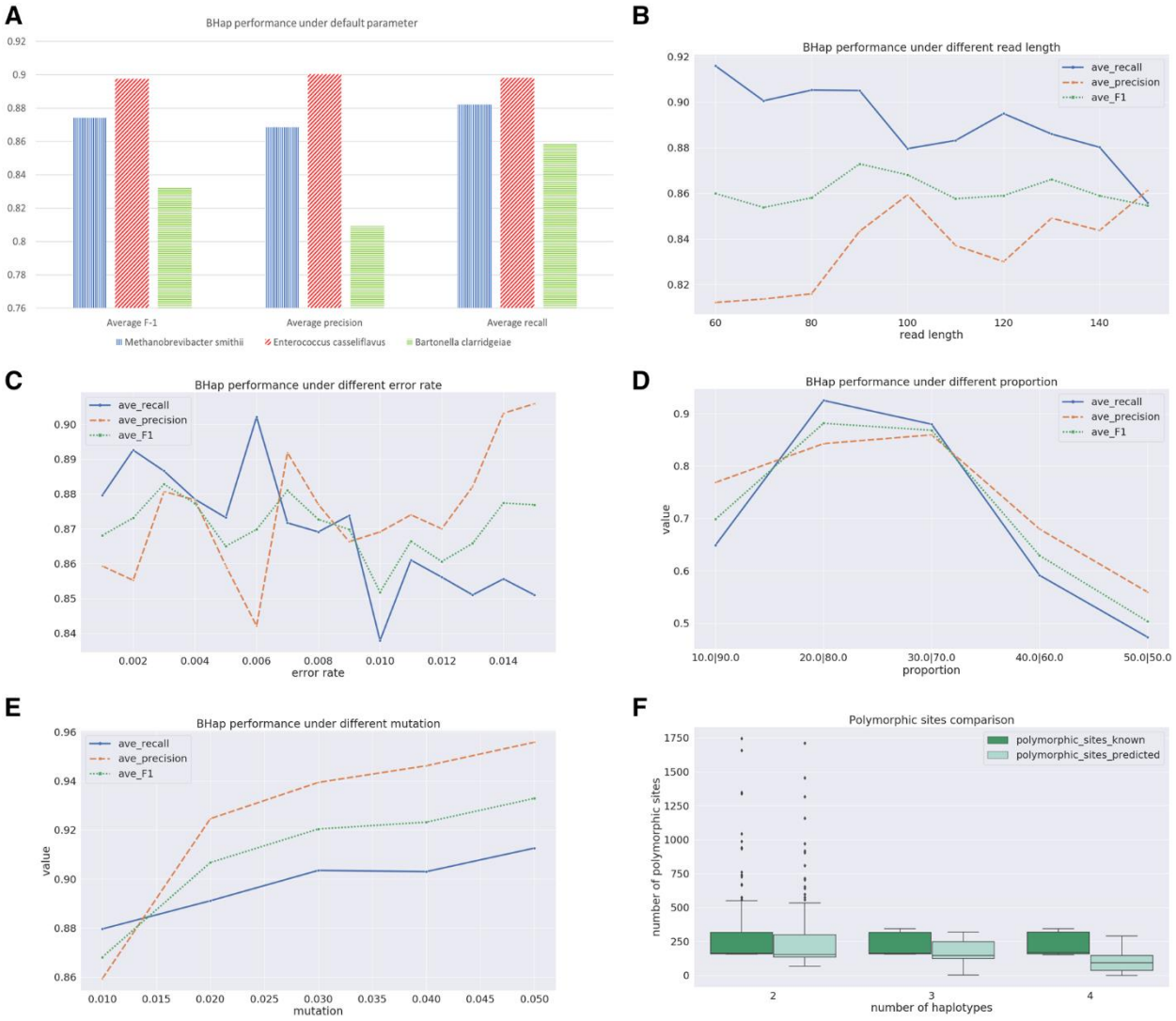

Figure 5 BHap performance under different parameters

(A) BHap performance under the default parameters. The three bars are for three different reference genomes; (B) BHap performance under different read lengths; (C) BHap performance under different error rates; (D) BHap performance under different strain proportions; (E) BHap performance under different mutation rates; (F) Predicted polymorphic sites by BHap compared with known polymorphic sites

Table 6 Performance under the default parameters for individual species

Species	F1	Precision	Recall
NC_009515.1	0.8527	0.8219	0.8864
NC_009515.1	0.8691	0.8723	0.8679
NC_009515.1	0.8962	0.8961	0.8986
NC_009515.1	0.8825	0.8889	0.8772
NC_009515.1	0.8707	0.8633	0.8802
NC_014932.1	0.836	0.8029	0.8725
NC_014932.1	0.7949	0.7748	0.8186
NC_014932.1	0.8572	0.8254	0.8916
NC_014932.1	0.8329	0.8203	0.8471
NC_014932.1	0.8419	0.8228	0.8631
NC_020995.1	0.9027	0.8908	0.9179
NC_020995.1	0.9021	0.8855	0.9209
NC_020995.1	0.9289	0.9404	0.9197
NC_020995.1	0.8389	0.8847	0.7992
NC_020995.1	0.9152	0.9001	0.9328
Average	0.8681	0.8593	0.8796

Table 7 Performance under different read lengths

Read length	Average F1	Average precision	Average recall
60	0.8599	0.8121	0.9159
70	0.8538	0.8136	0.9006
80	0.858	0.8159	0.9053
90	0.8729	0.8433	0.9051
100	0.8681	0.8593	0.8796
110	0.8576	0.8371	0.8832
120	0.859	0.83	0.895
130	0.8661	0.8491	0.886
140	0.8589	0.8437	0.8802
150	0.8546	0.8613	0.8559
Average	0.8609	0.8365	0.8907

Since sequencing errors may affect the polymorphism identification by BHap, we studied the BHap performance under different sequencing error rates (Figure 5C and Table 8). Compared with

the BHap performance under the default parameters, it fluctuated only slightly with the increment of sequencing error rates. For error rates from 0.1 to 1.5%, BHap had a minimum F1 score of 0.85, a minimum recall of 0.84 and a minimum precision of 0.84. On average, BHap had a F1 score of 0.87, a precision of 0.87 and a recall of 0.87. These numbers indicate that the BHap performance is quite robust to a variety of sequencing error rates.

Table 8 Performance under different sequence error rates

Error Rate	Average F1	Average precision	Average recall
0.001	0.8681	0.8593	0.8796
0.002	0.8731	0.8552	0.8926
0.003	0.8828	0.8807	0.8866
0.004	0.8773	0.8783	0.8785
0.005	0.865	0.8594	0.8732
0.006	0.8698	0.8421	0.9021
0.007	0.881	0.8919	0.8717
0.008	0.8727	0.877	0.8691
0.009	0.8698	0.8663	0.8738
0.01	0.8518	0.8691	0.8379
0.011	0.8664	0.874	0.861
0.012	0.8606	0.87	0.8561
0.013	0.8658	0.8821	0.851
0.014	0.8774	0.9032	0.8556
0.015	0.8769	0.9059	0.851
Average	0.8706	0.8743	0.8693

We also studied how the strain proportion affected the BHap performance (Figure 5D and Table 9). For instance, in a population with two strains, 10/90 or 50/50, with which proportion will BHap perform better? We observed that BHap performed the best at 20/80, where it had a F1 score of 0.88. We hypothesized that the proportion 10/90 may result in the best BHap performance, given a larger population coverage. We repeated the above experiments with larger coverage and BHap

indeed had better F1 scores with increased coverage (Table 9). The better performance of BHap with a 10/90 proportion was similar or better than that with the 20/80 proportion, suggesting that BHap performs differently on datasets with different strain proportions, and a higher coverage helps to reconstruct strains better.

Table 9 Performance under different proportions

Proportion	Coverage	Average F1	Average precision	average recall
10|90	100x	0.6984	0.7685	0.649
20|80	100x	0.8814	0.8426	0.925
30|70	100x	0.8681	0.8593	0.8796
40|60	100x	0.6291	0.6796	0.5913
50|50	100x	0.5034	0.5594	0.4734
10|90	200x	0.8709	0.8392	0.9065
10|90	300x	0.8856	0.8286	0.9535
10|90	400x	0.8852	0.8236	0.9597
10|90	500x	0.9098	0.8697	0.9557

We also investigated how mutation rates affected the performance of BHap (Figure 5E). We applied BHap to simulated datasets with different mutation rates (Table 10). BHap performed better with higher mutation rates. The F1 score increased by 0.065 when mutation rate increased from 0.01 to 0.05%. The F1 score with lower mutation rates was slightly decreased. This suggests that BHap can reconstruct strains with low mutation rates and is robust under different mutation rates.

Table 10 Performance under different mutation rates

Mutation Rate(%)	Average F1	Average precision	Average recall
0.01	0.8681	0.8593	0.8796
0.02	0.9067	0.9246	0.8911
0.03	0.9204	0.9394	0.9035
0.04	0.9232	0.9462	0.903
0.05	0.9329	0.9558	0.9126
Average	0.9103	0.9251	0.8980

It is worth mentioning that BHap accurately predicted strain proportions in almost all simulated datasets, including all datasets tested above (Table 11). In the first group of fifteen simulated datasets, with the known strain proportion of 30/70, the estimated strain proportions were 29.93/70.07, respectively. Even by changing the read length, sequence error rate and mutation rate, BHap robustly predicted an average strain proportion as 29.76/70.24, 29.37/70.63 and 29.85/70.15, respectively, for the default strain proportion 30/70. When the strain proportion was changed into 10/90, 20/80, 30/70, 40/60 and 50/50, the estimated proportion was 11.67/88.83, 19/81, 29.93/70.07, 34/66 and 33.33/66.67, respectively. When the strain proportion was changed into 10/50/60 and 10/20/30/40, the estimated proportion was 13.69/33.31/57.42 and 10.91/21.27/32.9/46.33, respectively. In summary, with the exception of the strain proportion 50/50, BHap reliably identified the strain proportions.

Table 11 Performance comparison between BHap and EVORhA under different proportions and coverage

Software	Proportion	Coverage	# of real strains	# of reconstructed strains proportion	Average F1	Average precision	Average recall	Reliability
BHap	30|70	50x	2	29/71	0.86	0.84	0.87	0.43
EVORhA	30|70	50x	2	33.67/66.33	0.18	0.50	0.11	0.09
BHap	30|70	100x	2	30.33/69.67	0.85	0.83	0.87	0.43
EVORhA	30|70	100x	2	30/70	0.23	0.51	0.15	0.11
BHap	30|70	150x	2	30/70	0.85	0.81	0.90	0.43
EVORhA	30|70	150x	2	30/70	0.25	0.51	0.17	0.12
BHap	30|70	200x	2	30/70	0.89	0.88	0.91	0.45
EVORhA	30|70	200x	2	31/69	0.26	0.49	0.19	0.13
BHap	10|30|60	50x	3	11/30/1959	0.59	0.59	0.59	0.28
EVORhA	10|30|60	50x	3	12/30.33/57.67	0.12	0.42	0.07	0.06
BHap	10|30|60	100x	3	13/31.67/55.33	0.68	0.63	0.74	0.34
EVORhA	10|30|60	100x	3	11/26.67/62.33	0.17	0.44	0.11	0.09
BHap	10|30|60	150x	3	10.33/32/57.67	0.79	0.72	0.87	0.39
EVORhA	10|30|60	150x	3	11.33/28/60.67	0.19	0.37	0.13	0.09
BHap	10|30|60	200x	3	9.67/32/58.33	0.84	0.79	0.91	0.42
EVORhA	10|30|60	200x	3	10.67/29.67/59.67	0.20	0.36	0.14	0.10
BHap	10|20|30|40	50x	4	0/20.67/37.67/41.67	0.32	0.27	0.45	0.16
EVORhA	10|20|30|40	50x	4	11/21/29.67/38.33	0.10	0.33	0.06	0.05
BHap	10|20|30|40	100x	4	9/17/33.33/43.67	0.36	0.29	0.5	0.18
EVORhA	10|20|30|40	100x	4	10/18.33/29.33/42.33	0.14	0.34	0.091	0.07
BHap	10|20|30|40	150x	4	7.33/14.33/31.67/46.67	0.28	0.22	0.42	0.14
EVORhA	10|20|30|40	150x	4	10.33/16/33/40.67	0.15	0.29	0.11	0.08
BHap	10|20|30|40	200x	4	7/13/31.33/48.67	0.37	0.36	0.54	0.19
EVORhA	10|20|30|40	200x	4	10.33/18.33/33.67/40.67	0.17	0.29	0.12	0.08

We also want to point out that BHap predicted the number of polymorphic sites reasonably well in simulated datasets (Figure 5F), especially in datasets with two strains. In the above simulated datasets, with two strains in a population, BHap had a recall of 0.86, with a standard deviation of 0.11. With three strains in a population, the recall was 0.78, with a standard deviation of 0.22. With four strains in a population, the recall became 0.54, with a standard deviation of 0.32.

Correspondingly, the reliability score in these three scenarios was, 0.43, 0.36 and 0.17, respectively. Note that the largest reliability in theory is 0.50.

The lower recall and reliability above for three and four strains may be due to the relatively small coverage in these datasets, most of which has a coverage of 100×. We hypothesized that the recall and the reliability were also reasonably good for bacterial populations with more than two strains, given a higher sequencing depth. We thus examined the recall and the reliability when the coverage was high (Table 11). We found that when coverage increased, the recall and the reliability increased as well. For the coverage of 500×, the recall was 0.92, 0.85 and 0.77, and the reliability was 0.47, 0.42 and 0.28 for two, three and four strains, respectively. This implied that BHap is able of reliably predict polymorphic sites in bacterial populations, given a high sequencing depth.

3.3.2 BHap reconstructs strains better than EVORhA on simulated datasets

Since EVORhA is the first and the only existing strain reconstruction tool for bacterial populations, we compared BHap with EVORhA on 216 simulated datasets (the 7th–9th and 10th–11th groups in Table 5), with three bacterial species, different numbers of strains (2–4), different sequencing depth (50×, 100×, 150×, 200× and 500×), different mutation rates (0.01, 0.02, 0.05, 0.07, 0.1 and 0.15%) and different evolution trajectories (no evolution relation, T0, T1 and T2).

On 36 datasets in the seventh group, on average, BHap had a F1 score of 0.64 while EVORhA had a F1 score of 0.18 (Table 11 and Table 12). For populations with two strains, the average F1 score of BHap was 0.86 and the average F1 score of EVORhA was 0.23. For populations with three strains, BHap had an average F1 score of 0.72 while EVORhA had 0.17. For populations with four strains, BHap performed better than EVORhA as well (F1 score of 0.33 versus 0.14). In terms of

different population coverage, given a strain proportion, the higher the coverage was, the higher F1 scores was for both BHap and EVORhA (Table 12).

Table 12 Performance comparison of BHap with EVORhA on the seventh group of simulated datasets

Proportion(Coverage)	# of reconstructed strains	Average F1	Average precision	Average recall
30|70(50x)	2.33(4.33)	0.86(0.18)	0.84(0.50)	0.87(0.11)
30|70(100x)	3(4.67)	0.85(0.23)	0.83(0.51)	0.87(0.15)
30|70(150x)	3.67(5.67)	0.85(0.25)	0.81(0.51)	0.90(0.17)
30|70(200x)	3.33(5.33)	0.89(0.26)	0.88(0.49)	0.91(0.19)
10|30|60(50x)	3.0(5.33)	0.59(0.12)	0.59(0.42)	0.59(0.07)
10|30|60(100x)	4.0(5.67)	0.68(0.17)	0.63(0.44)	0.74(0.11)
10|30|60(150x)	4.0(6.7)	0.79(0.19)	0.72(0.37)	0.87(0.13)
10|30|60(200x)	4.0(5.33)	0.84(0.20)	0.79(0.36)	0.91(0.14)
10|20|30|40(50x)	2.67(5.67)	0.32(0.10)	0.27(0.33)	0.45(0.06)
10|20|30|40(100x)	4.0(7.33)	0.36(0.14)	0.29(0.34)	0.5(0.09)
10|20|30|40(150x)	5.0(6.0)	0.28(0.15)	0.22(0.29)	0.42(0.11)
10|20|30|40(200x)	5.33(6.0)	0.37(0.17)	0.36(0.29)	0.54(0.12)

Note: In the last four columns, the first number is for BHap and the number in the parenthesis is for EVORhA.

We also noticed that EVORhA produced many false positive strains per dataset, especially when there were more strains in populations (Table 12). For 12 datasets with two strains, on average, BHap predicted 3.08 strains per samples while EVORhA predicted five strains per samples. For another two groups of twelve datasets with three and four strains, respectively, BHap predicted 3.75 and 4.25 strains while EVORhA predicted 5.75 and 6.25 strains, respectively. EVORhA may know that it predicted much more strains than actual ones, but it did not provide a way to filter the false positive ones.

Since EVORhA had the best performance at 500× coverage and higher mutation rates (Pulido-Tamayo et al., 2015), we further compared BHap with EVORhA on the eighth group of eighteen datasets with 500× coverage (Table 13). BHap had an average F1 score of 0.78, a precision of 0.75 and a recall of 0.84. Correspondingly, EVORhA only had an average F1 score of 0.21, a precision of 0.38 and a recall of 0.15 (Table 13). The low recall from EVORhA suggested that it may be not good at predicting actual polymorphisms in the reconstructed strains. We also compared EVORhA with BHap on the ninth group of 45 datasets with very high mutation rates (0.07, 0.1 and 0.15%) (Table 14). On these datasets, BHap had an average F1 score of 0.94, a precision of 0.96 and a recall of 0.92, while EVORhA had 0.12, 0.48 and 0.07, respectively. The performance of BHap was significantly improved with higher mutation rates, while EVORhA still had a low recall and F1 score.

Table 13 Performance comparison between BHap and EVORhA under high coverage and high mutation rates

Software	Coverage Info	Coverage	# of strain	mutation rate (%)	predicted proportion	average F-1	Average precision	average recall	Reliability
BHap	30|70	500x	2	0.02	30.1|69.9	0.8897	0.9011	0.8817	0.4448
EVORhA	30|70	500x	2	0.02	30.47|69.53	0.2795	0.4994	0.1962	0.1398
BHap	30|70	500x	2	0.05	30.05|69.95	0.9292	0.9618	0.9013	0.4646
EVORhA	30|70	500x	2	0.05	28.69|71.31	0.2163	0.507	0.1398	0.1081
BHap	10|30|60	500x	3	0.02	11.16|36.73|52.12	0.8267	0.8044	0.8525	0.4133
EVORhA	10|30|60	500x	3	0.02	11.05|29.1|59.86	0.222	0.4133	0.1594	0.111
BHap	10|30|60	500x	3	0.05	12.07|39.59|48.34	0.8622	0.8725	0.8547	0.4311
EVORhA	10|30|60	500x	3	0.05	10.19|29.89|59.92	0.2163	0.3334	0.163	0.1082
BHap	10|20|30|40	500x	4	0.02	7.58|15.57|23.01|53.84	0.5718	0.4787	0.7882	0.2859
EVORhA	10|20|30|40	500x	4	0.02	9.25|18.22|29.72|42.82	0.1762	0.2702	0.134	0.0881
BHap	10|20|30|40	500x	4	0.05	8.08|16.4|24.14|51.39	0.575	0.4995	0.7561	0.2875
EVORhA	10|20|30|40	500x	4	0.05	10.34|16.57|31.17|41.91	0.156	0.2416	0.1179	0.078

Table 14 Performance comparison between BHap and EVORhA under high mutation rates

Software	Coverage Info	mutation rate (%)	predicted proportion	average F-1	Average precision	average recall	Reliability
BHap	30|70	0.07	29.28|70.72	0.9166	0.9476	0.8879	0.4583
EVORhA	30|70	0.07	34.71|65.29	0.0721	0.4946	0.0392	0.036
BHap	30|70	0.07	29.76|70.24	0.9382	0.9614	0.9173	0.4691
EVORhA	30|70	0.07	32.73|67.27	0.1007	0.4921	0.0564	0.0503
BHap	30|70	0.07	29.94|70.06	0.9434	0.9617	0.9268	0.4717
EVORhA	30|70	0.07	31.85|68.15	0.1092	0.4631	0.063	0.0546
BHap	30|70	0.07	29.9|70.1	0.9393	0.9525	0.9273	0.4696
EVORhA	30|70	0.07	36.59|63.41	0.1468	0.4464	0.0893	0.0734
BHap	30|70	0.07	30.01|69.99	0.933	0.964	0.9058	0.4665
EVORhA	30|70	0.07	30.48|69.52	0.2416	0.5032	0.1603	0.1208
BHap	30|70	0.1	29.4|70.6	0.9162	0.9507	0.8845	0.4581
EVORhA	30|70	0.1	32.8|67.2	0.0507	0.4596	0.027	0.0254
BHap	30|70	0.1	29.7|70.3	0.9357	0.9588	0.9149	0.4679
EVORhA	30|70	0.1	35.17|64.83	0.0949	0.5063	0.0531	0.0474
BHap	30|70	0.1	29.82|70.18	0.9487	0.9697	0.9294	0.4743
EVORhA	30|70	0.1	35.93|64.07	0.1101	0.4451	0.0638	0.055
BHap	30|70	0.1	29.85|70.15	0.9513	0.9692	0.9348	0.4756
EVORhA	30|70	0.1	31.67|68.33	0.1355	0.4961	0.08	0.0677
BHap	30|70	0.1	30.0|70.0	0.9366	0.9669	0.91	0.4683
EVORhA	30|70	0.1	30.46|69.54	0.2442	0.4742	0.1662	0.1221
BHap	30|70	0.15	29.6|70.4	0.9157	0.9616	0.8743	0.4579
EVORhA	30|70	0.15	34.65|65.35	0.0417	0.4784	0.0219	0.0208
BHap	30|70	0.15	29.51|70.49	0.9486	0.9705	0.928	0.4743
EVORhA	30|70	0.15	32.22|67.78	0.0644	0.5108	0.0346	0.0322
BHap	30|70	0.15	29.7|70.3	0.9516	0.9733	0.9315	0.4758
EVORhA	30|70	0.15	38.53|61.47	0.0788	0.4167	0.044	0.0394
BHap	30|70	0.15	29.84|70.16	0.9522	0.9697	0.936	0.4761
EVORhA	30|70	0.15	32.92|67.08	0.1051	0.4868	0.06	0.0525
BHap	30|70	0.15	29.97|70.03	0.9462	0.9723	0.9223	0.4731
EVORhA	30|70	0.15	29.92|70.08	0.2154	0.4999	0.1403	0.1077

We also compared BHap with EVORhA on the 117 datasets where strains had specified evolutionary relationship (the 10th and 11th groups, Tables 15 and 16). In every case, BHap had a larger F1 score, precision and recall than EVORhA. On the 10th group of 72 datasets, we studied

how BHap and EVORhA performed under different coverage and strain proportions. BHap had an average F1 score of 0.71 (0.49), a precision of 0.78 (0.50) and a recall of 0.66 (0.52) on datasets with three (four) strains. Correspondingly, EVORhA had a F1 score of 0.36 (0.28), a precision of 0.45 (0.29), a recall of 0.33 (0.28) on datasets with three (four) strains. In different datasets, the performance of BHap consistently changed with that of EVORhA, in the sense that when BHap had a better performance, EVORhA had a better performance, and vice versa. Both BHap and EVORhA estimated the strain proportions relatively well, especially when coverage was high. However, the higher coverage did not always result in better performance for EVORhA and BHap, although the F1 score was often the highest at the coverage 400× or 500×. Higher coverage may not result in better F1 scores, because the reads were not necessarily evenly distributed and the complexity in terms of sharing polymorphic sites by strains, which may result in different number of predicted strains and thus different accuracy. Since EVORhA performed better with higher mutation rates, we further compared EVORhA with BHap on the 11th group of 45 datasets, in which we studied how EVORhA and BHap performed with different mutation rates. Consistent with the above study, both tools performed better with higher mutation rates (Table 16).

Table 15 Performance comparison between BHap and EVORhA under different coverage, mutation rates, and evolution trajectories

Software	similarity type	real proportion	coverage	predicted proportion	average reliability	average F1	average precision	average recall
BHap	T0	10|30|60	200x	9.62|33.36|57.03	0.366	0.732	0.775	0.702
EVORhA	T0	10|30|60	200x	10.96|31.15|57.89	0.145	0.29	0.371	0.244
BHap	T0	10|30|60	300x	9.34|33.47|57.19	0.366	0.733	0.768	0.712
EVORhA	T0	10|30|60	300x	10.54|30.08|59.38	0.212	0.424	0.443	0.46
BHap	T0	10|30|60	400x	7.72|35.01|57.26	0.334	0.667	0.751	0.61
EVORhA	T0	10|30|60	400x	9.82|29.91|60.28	0.189	0.379	0.365	0.398
BHap	T0	10|30|60	500x	12.19|39.35|48.46	0.32	0.639	0.786	0.555
EVORhA	T0	10|30|60	500x	10.01|30.47|59.51	0.172	0.344	0.365	0.332
BHap	T0	5|25|70	200x	6.23|26.44|67.33	0.349	0.699	0.775	0.64
EVORhA	T0	5|25|70	200x	5.17|25.53|69.29	0.164	0.329	0.533	0.257
BHap	T0	5|25|70	300x	5.37|26.8|67.83	0.363	0.726	0.773	0.687
EVORhA	T0	5|25|70	300x	6.06|25.42|68.51	0.132	0.264	0.531	0.193
BHap	T0	5|25|70	400x	4.98|26.96|68.07	0.378	0.755	0.828	0.7
EVORhA	T0	5|25|70	400x	5.34|24.92|69.74	0.217	0.434	0.555	0.398
BHap	T0	5|25|70	500x	4.83|27.0|68.17	0.358	0.717	0.79	0.665
EVORhA	T0	5|25|70	500x	4.88|25.38|69.74	0.191	0.382	0.425	0.356
BHap	T0 average	None	None	None	0.354	0.708	0.781	0.659
EVORhA	T0 average	None	None	None	0.178	0.356	0.449	0.33
BHap	T1	5|15|25|55	200x	5.54|12.44|29.02|53.0	0.222	0.443	0.447	0.475
EVORhA	T1	5|15|25|55	200x	6.76|14.9|25.63|52.71	0.109	0.217	0.265	0.199
BHap	T1	5|15|25|55	300x	3.73|10.53|30.78|54.96	0.221	0.443	0.412	0.532
EVORhA	T1	5|15|25|55	300x	4.7|15.24|24.38|55.68	0.132	0.263	0.242	0.297
BHap	T1	5|15|25|55	400x	3.81|11.22|29.35|55.62	0.242	0.484	0.456	0.576
EVORhA	T1	5|15|25|55	400x	4.53|15.44|24.54|55.49	0.161	0.323	0.308	0.347
BHap	T1	5|15|25|55	500x	5.48|16.8|25.73|51.99	0.268	0.535	0.564	0.585
EVORhA	T1	5|15|25|55	500x	5.01|15.16|24.89|54.94	0.159	0.317	0.267	0.401
BHap	T1	5|10|35|50	200x	5.13|19.79|27.79|47.29	0.192	0.385	0.415	0.403

Software	similarity type	real proportion	coverage	predicted proportion	average reliability	average F1	average precision	average recall
EVORhA	T1	5|10|35|50	200x	6.24|10.67|28.58|54.5	0.098	0.197	0.271	0.161
BHap	T1	5|10|35|50	300x	5.96|17.71|27.35|48.99	0.213	0.427	0.464	0.448
EVORhA	T1	5|10|35|50	300x	4.5|7.7|34.3|53.5	0.114	0.228	0.246	0.226
BHap	T1	5|10|35|50	400x	5.59|9.66|27.73|57.02	0.215	0.429	0.432	0.465
EVORhA	T1	5|10|35|50	400x	4.43|9.11|35.24|51.22	0.141	0.283	0.315	0.27
BHap	T1	5|10|35|50	500x	5.25|10.07|30.87|53.82	0.241	0.482	0.5	0.51
EVORhA	T1	5|10|35|50	500x	5.27|10.96|35.16|48.61	0.146	0.292	0.336	0.281
BHap	T1 average	None	None	None	0.227	0.453	0.461	0.499
EVORhA	T1 average	None	None	None	0.133	0.265	0.281	0.27
BHap	T2	5|15|25|55	200x	5.79|11.4|29.32|53.49	0.227	0.454	0.443	0.5
EVORhA	T2	5|15|25|55	200x	5.59|16.66|24.29|53.46	0.106	0.212	0.236	0.201
BHap	T2	5|15|25|55	300x	4.96|15.22|25.85|53.97	0.255	0.51	0.527	0.555
EVORhA	T2	5|15|25|55	300x	4.85|16.07|24.41|54.67	0.133	0.266	0.254	0.286
BHap	T2	5|15|25|55	400x	5.52|18.34|24.47|51.67	0.284	0.567	0.598	0.611
EVORhA	T2	5|15|25|55	400x	4.63|15.27|23.88|56.22	0.144	0.288	0.263	0.332
BHap	T2	5|15|25|55	500x	5.54|18.13|25.18|51.15	0.295	0.59	0.624	0.621
EVORhA	T2	5|15|25|55	500x	4.83|13.46|21.35|60.36	0.168	0.336	0.326	0.369
BHap	T2	5|10|35|50	200x	6.05|19.22|26.96|47.77	0.248	0.495	0.512	0.518
EVORhA	T2	5|10|35|50	200x	5.12|9.8|30.56|54.52	0.11	0.221	0.275	0.195
BHap	T2	5|10|35|50	300x	6.78|15.26|30.31|47.64	0.248	0.496	0.515	0.519
EVORhA	T2	5|10|35|50	300x	4.95|10.1|27.83|57.13	0.144	0.289	0.326	0.272
BHap	T2	5|10|35|50	400x	5.75|11.15|30.3|52.8	0.241	0.482	0.484	0.514
EVORhA	T2	5|10|35|50	400x	4.19|8.23|31.82|55.76	0.161	0.321	0.351	0.317
BHap	T2	5|10|35|50	500x	7.4|15.18|26.87|50.56	0.273	0.546	0.587	0.554
EVORhA	T2	5|10|35|50	500x	4.21|8.86|30.58|56.35	0.169	0.339	0.331	0.354
BHap	T2 average	None	None	None	0.259	0.518	0.536	0.549

Software	similarity type	real proportion	coverage	predicted proportion	average reliability	average F1	average precision	average recall
EVORhA	T2 average	None	None	None	0.142	0.284	0.295	0.291

Three types of evolution trajectories are considered in this table (the first column). In the first type, the T0 type, there are three strains in a population, with the first two strains sharing 30% of their polymorphic sites and the third strain sharing no polymorphic site with the first two strains. In the second type, the T1 type, there are four strains in a population, with the first two strains sharing 30% of their polymorphic sites, the third strain sharing 10% of polymorphic sites with the first two, and the fourth strain sharing no polymorphic site with the first three. In the third type, the T2 type, there are four strains in a population, with the first two strains sharing 30% of their polymorphic sites, the third and the fourth strains sharing 30% of their polymorphic sites, and the two pairs of strains sharing no polymorphic site. The strain proportion in the second column gives the corresponding proportion for each strain in the population, in the order of the first, second, third, and fourth if there are four strains in a population. There are three datasets corresponding to the three reference genomes used for the numbers in each row. The row with the name "ti average" is the average performance of the rows with the i-th strain similarity type, where i is 0, 1, or 2.

Table 16 Performance comparison between BHap and EVORhA under different mutation rates

similarity type	mutation rate (%)	real proportion	coverage	predicted proportion	average reliability	average F1	average precision	average recall
T0	0.01	5|25|70	300x	5.37|26.8|67.83(5.23|25.05|69.72)	0.363(0.131)	0.725(0.262)	0.773(0.517)	0.686(0.185)
T0	0.02	5|25|70	300x	5.4|26.46|68.14(5.17|25.53|69.29)	0.377(0.164)	0.755(0.329)	0.812(0.533)	0.707(0.257)
T0	0.03	5|25|70	300x	5.37|26.39|68.24(5.37|25.45|69.18)	0.374(0.12)	0.749(0.24)	0.839(0.466)	0.677(0.167)
T0	0.04	5|25|70	300x	5.4|26.64|67.96(5.22|25.52|69.26)	0.376(0.116)	0.752(0.233)	0.84(0.468)	0.682(0.161)
T0	0.05	5|25|70	300x	5.44|26.45|68.11(5.42|24.92|69.65)	0.38(0.115)	0.76(0.23)	0.85(0.417)	0.688(0.161)
T0	None	None	None	None	0.374(0.129)	0.748(0.259)	0.823(0.48)	0.688(0.186)
T1	0.01	5|10|35|50	300x	5.07|17.07|26.91|50.96(5.13|8.39|37.02|49.46)	0.211(0.102)	0.422(0.203)	0.441(0.284)	0.457(0.164)
T1	0.02	5|10|35|50	300x	5.35|15.97|24.9|53.77(4.68|9.68|35.63|50.0)	0.219(0.071)	0.438(0.143)	0.481(0.228)	0.455(0.106)
T1	0.03	5|10|35|50	300x	5.01|14.39|30.9|49.7(4.99|8.34|36.14|50.52)	0.213(0.091)	0.426(0.182)	0.442(0.287)	0.445(0.139)
T1	0.04	5|10|35|50	300x	5.32|15.84|24.16|54.67(4.5|7.7|34.3|53.5)	0.216(0.114)	0.433(0.228)	0.455(0.246)	0.443(0.226)
T1	0.05	5|10|35|50	300x	5.18|15.27|26.85|52.71(5.63|9.51|33.16|51.7)	0.227(0.085)	0.453(0.169)	0.468(0.277)	0.468(0.127)
T1	None	None	None	None	0.217(0.093)	0.434(0.185)	0.458(0.264)	0.454(0.153)
T2	0.01	5|10|35|50	300x	6.96|15.56|30.64|46.84(4.95|10.1|27.83|57.13)	0.25(0.144)	0.501(0.289)	0.513(0.326)	0.526(0.272)
T2	0.02	5|10|35|50	300x	5.94|17.38|26.93|49.75(4.74|8.36|35.9|51.0)	0.261(0.093)	0.522(0.187)	0.548(0.298)	0.532(0.142)
T2	0.03	5|10|35|50	300x	5.57|17.41|28.57|48.45(4.61|9.9|34.5|50.99)	0.247(0.105)	0.495(0.209)	0.543(0.299)	0.5(0.165)
T2	0.04	5|10|35|50	300x	5.66|17.66|29.64|47.04(4.62|9.73|32.81|52.84)	0.239(0.122)	0.479(0.245)	0.523(0.315)	0.484(0.205)
T2	0.05	5|10|35|50	300x	5.44|16.59|26.84|51.13(5.06|9.34|35.32|50.28)	0.242(0.094)	0.483(0.188)	0.513(0.323)	0.491(0.136)
T2	None	None	None	None	0.248(0.112)	0.496(0.224)	0.528(0.312)	0.507(0.184)

Three types of evolution trajectories are considered in this table (the first column). In the first type, the T0 type, there are three strains in a population, with the first two strains sharing 30% of their polymorphic sites and the third strain sharing no polymorphic site with the first two strains. In the second type, the T1 type, there are four strains in a population, with the first two strains sharing 30% of their polymorphic sites, the third strain sharing 10% of polymorphic sites with the first two, and the fourth strain sharing no polymorphic site with the first three. In the third type, the T2 type, there are four strains in a population, with the first two strains sharing 30% of their polymorphic sites, the third and the fourth strains sharing 30% of their polymorphic sites, and the two pairs of strains sharing no polymorphic site. The strain proportion in the second column gives the corresponding proportion for each strain in the population, in the order

of the first, second, third, and fourth if there are four strains in a population. There are three datasets corresponding to the three reference genomes used for the numbers in each row. The row with the name "ti average" is the average performance of the rows with the i-th strain similarity type, where i is 0, 1, or 2.

3.3.3 BHap reconstructs strains better than EVORhA on experimental dataset

We compared BHap and EVORhA on two experimental datasets (Section 2). With strains unknown in these datasets, we could not calculate F1 score, precision and recall. We thus focused on comparing their reliability in two ways. One was strain based, where the most similar pairs of strains were predicted by a tool, with one from a population and mixed sample, and the other from its corresponding clone or unmixed sample, and then reliability was calculated based on these pairs of strains. The other was SAMtools based, where we compared the polymorphic sites in the predicted strains in the population or mixed sample by the tool with the polymorphic sites from raw reads for the corresponding clone or unmixed sample inferred by SAMtools directly. We found that BHap had a higher reliability than EVORhA based on both approaches.

By the strain based approach, BHap had an average reliability of 0.09 and 0.10 on the mixed infection dataset and the evolved population dataset, respectively, while EVORhA had an average reliability of 0.01 and 0.03, correspondingly (Figure 6, Table 17, Supplementary Table S2 in original paper) (Xin Li et al., 2019). The reliability of BHap was significantly larger than that of EVORhA on the mixed infection dataset (Mann-Whitney test p-value 3.35 × 10-6). We did not consider the significance of the reliability difference on the evolved dataset, as there were only three time points involved. By the SAMtools based approach, BHap had an average reliability of 0.09 and 0.09 on the mixed infection dataset and the evolved population dataset, respectively, while EVORhA correspondingly had 0.01 and 0.01, respectively (Figure 6, Table 17, Supplementary Table S2 in original paper) (Xin Li et al., 2019). The reliability of BHap was

significantly larger than that of EVORhA on the mixed infection dataset (Mann-Whitney test p-value 3.347×10^{-6}).

One should focus on the relative reliabilities above. The actual reliability of both tools seemed not large. In theory, the largest reliability is 0.50, when polymorphic sites from the clone or the unmixed sample are exactly the same as those from the corresponding population. Polymorphic sites can be added or removed in practice, making the reliability lower. In fact, we applied SAMtools to the 54 mixed infection datasets and their corresponding unmixed datasets to define polymorphic sites directly and calculated the reliability, the reliability was from 0.09 to 0.47, with a mean of 0.22 and a median of 0.20. Such a reliability was based on the assumption that there was only one strain in the population and the same one in the clones or unmixed samples. Since there may be different strains in clones, unmixed samples and population, different strains and their pairing most likely result in much smaller reliability. It is thus likely that we may have achieved the best reliabilities on these experimental datasets. More importantly, BHap had much higher reliability than EVORhA.

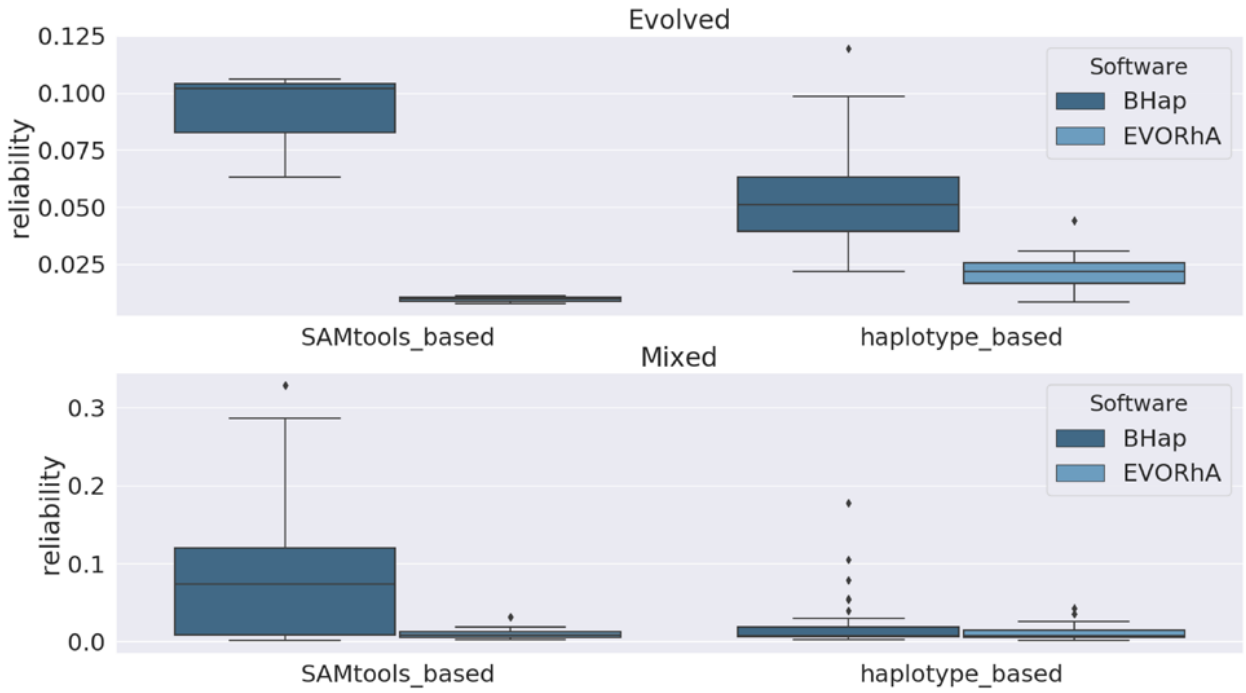

Figure 6 Reliability comparison on experimental datasets

The box plot for BHap is in front of that for EVORhA in the four comparisons

Table 17 Performance comparison between BHap and EVORhA under evolved population dataset

clone name	population name	time point	reliability(BHap(EVORhA))	
			strain based	SAMtools based
SRR1585349	SRR1585342	TP0	0.093(0.024)	0.102(0.011)
SRR1585353	SRR1585343	TP1	0.12(0.044)	0.063(0.008)
SRR1585354	SRR1585346	TP2	0.08(0.031)	0.106(0.01)

In the last two columns, the number(s) in the parenthesis is for EVORhA while the number(s) not in the parenthesis is for BHap. We evaluated BHap and EVORhA in two ways: (1) Strain based method: we apply a tool to predict strains in the population sample and in its corresponding clone sample. We then find one pair of strains with the largest percentage of shared polymorphic sites, with one strain from the population sample and the other from its clone sample. We then calculate the reliability based on this pair of strains and obtain the reliability of the tool; (2). SAMtools based: we apply a tool to predict strains in the population sample. We apply SAMTools to identify polymorphic sites in the corresponding clone sample. We then compare the polymorphic sites in the predicted strains from the population sample with those from the corresponding clone sample. In this way, we also have a pairs of polymorphic site sets. We then calculate the reliability similarly as above.

3.4 Discussion

We developed a novel strain reconstruction method for bacterial populations, called BHap. With an estimated global view of the coverage of strains, BHap decomposes flows of polymorphic sites in the network and finds a set of feasible flows, each representing a strain. BHap repeats this process with different k-mer lengths and combines the results from different k-mer lengths to generate robust predictions. Such a global view based approach may prevent from the expansion of the errors made at local polymorphic sites and avoid the difficulty in extending these local sites based on the adjacent local sites. Tested on simulated datasets, BHap shows a high F1 score, precision and recall. Compared with EVORhA, BHap shows much better accuracy in terms of F1 score, precision, recall and reliability.

In addition to EVORhA, we also attempted to compare BHap with ShoRAH (Zagordi et al., 2011), one of the most highly cited strain reconstruction tools for viral populations. We were unable to run it on our simulated datasets. Neither were we able to test it on the two experimental datasets. The tool could not be run, likely because of the much lower mutation rates in these datasets. In fact, the EVORhA study also mentioned that ShoRAH cannot be run on bacterial genome (Pulido-Tamayo et al., 2015).

For the mixed infection dataset, the proportion of two unmixed samples in the corresponding mixed sample was provided in Supplementary S2 of original paper (Xin Li et al., 2019). We tried to compare the predicted proportions with the known ones in these datasets and found that they often did not agree well. By further applying SAMtools to every mixed sample and every unmixed sample and then comparing the identified polymorphic sites from two samples, we noticed that the correspondence between the unmixed samples and the mixed samples provided in the above link could be wrong. For instance, we often found another pair of unmixed samples had more polymorphic sites shared with a mixed sample than its assigned pair of unmixed samples in

Supplementary S2 of original paper (Xin Li et al., 2019). Therefore, we believed that the correspondence of samples and their proportions provided in this link may be inaccurate.

We compared the predicted polymorphic sites by BHap with the 'known' polymorphic sites. BHap is able to predict the known polymorphic sites in simulated datasets, as shown in Figure 5F. However, it predicts much fewer 'known' polymorphic sites in experimental datasets, although it predicts much more 'known' sites than EVORhA. This is likely that the 'known' polymorphic sites in the experimental datasets cannot be well defined. It may also suggest that there is still room for the improvement of the bacterial strain reconstruction methods and tools.

BHap depends on the coverage difference of strains in a population to distinguish these strains. Our study shows that it works well on datasets with different coverage of strains, such as a 30/70 strain proportion. However, it has a much lower F1 score around 0.50 when the strains have the same coverage based on our study on simulated datasets. In this regard, BHap is not applicable to every bacterial population. Moreover, although BHap performs well in terms of identifying known polymorphic sites, it does not precisely identify all known polymorphic sites. This may be caused by the de Bruijn graph data structures and related Velvet libraries we used. In addition, we also noticed that BHap did not perform so well in populations with a specified evolution trajectory as in populations with no shared polymorphic sites among strains, implying that it is important to take the evolution information into account for the inference and prediction. In the future, we hope to further improve BHap by taking these aspects into account.

CHAPTER FOUR: A NOVEL TOOL FOR BACTERIAL STRAIN RECONSTRUCTION FROM READS

Previously published as Li, X., Hu, H., & Li, X. (2020). mixtureS: a novel tool for bacterial strain reconstruction from reads. Bioinformatics.

4.1 Introduction

It is imperative to reconstruct bacterial strain genomes from shotgun reads of clonal samples of individual species or metagenomic samples of many species (Luo et al., 2015; Pulido-Tamayo et al., 2015). Bacterial genomes are constantly evolving, where mutations are accumulated in different copies of a species genome that result in different strain genomes of the same species mixed in a sample (Zolfo, Tett, Jousson, Donati, & Segata, 2017). To identify bacterial strains in a sample, shotgun sequencing is routinely employed to generate short DNA segments from mixed strain genomes in a sample, which are called reads and approximate the full DNA content and abundance of the mixed strain genomes in the sample (Nayfach, Rodriguez-Mueller, Garud, & Pollard, 2016). To reconstruct the strain genomes from these reads is thus crucial for our understanding of the bacterial diversity, evolution, function, drug resistance, etc. (Nayfach et al., 2016; Pulido-Tamayo et al., 2015; Truong, Tett, Pasolli, Huttenhower, & Segata, 2017; Zolfo et al., 2017). More than a dozen methods are available for strain studies. The vast majority of them depend on known strains and/or known variations in strains, or intent to identify only variations in the species genome or a portion of the strain genomes, which cannot be generally applied or cannot de novo reconstruct the entire strain genomes (Ahn, Chai, & Pan, 2015; Albanese & Donati, 2017; Hong et al., 2014; Luo et al., 2015; Nayfach et al., 2016; Quince et al., 2017; Roosaare et al., 2017; Truong et al., 2017; Zolfo et al., 2017). This leaves only a few methods that can de novo reconstruct

bacterial strain genomes from reads in individual samples (Xin Li et al., 2019; Pulido-Tamayo et al., 2015; Smillie et al., 2018). Moreover, to our knowledge, the performance of these remaining methods is still suboptimal. In addition, some tools are difficult to use by general biologists. We thus create a new tool called mixtureS that have better accuracy and are more user-friendly.

4.2 Materials and Methods

4.2.1 Simulated datasets

There are 243 simulated datasets with different configurations, including different coverage, sequence error rates, mutation rates, strain proportions and numbers of strains (http://www.cs.ucf.edu/~xiaoman/mixtureS/simulated243). Coverage is the sequencing depth of a dataset, which is equal to the ratio of the sum of the length of all reads to the length of the reference genome. The sequence error in next-generation sequencing (NGS) is self-explanation. The mutation rate is the percentage of the variations in a strain genome compared with its corresponding species reference genome. The strain proportion is the relative abundance of reads from different strains. For example, a strain proportion in a dataset of 10/30/60 means that reads from the first, second and third strain account for 10%, 30% and 60%, respectively, of the total reads in this dataset. Each simulated dataset contains 2, 3 or 4 strains from a bacterial species in this study. For each configuration, the same number of datasets are generated for each of the three randomly selected bacterial species: Bartonella clarridgeiae, Enterococcus casseliflavus and Methanobrevibacter smithii (GenBank NC_014932, NC_020995, and NC_009515, respectively). The dwgsim tool (https://github.com/nh13/DWGSIM) was used to simulate short 100 bp long

paired-end Illumina reads for each strain. All simulated reads from all strains specified in a dataset were then mixed together to produce a simulated dataset.

We simulated eight groups of simulated datasets, and the datasets are stored in the link (http://www.cs.ucf.edu/~xiaoman/mixtureS/simulated243). The first group was the default group with 3 datasets. Here were the default values of the simulation parameters: mutation rate 0.01%; strain proportion 30/70; sequence error rate 0.1%; and coverage 200x. For the remaining seven groups, one or more parameters were changed. The second group with 12 datasets was to study the effect of different mutation rate, in which the mutate rate was changed from 0.02% to 0.05%. The third group contained 12 datasets under different strain proportions (40/60, 50/50, 20/80, 10/90). The fourth group had 42 datasets with different sequenced error rates (from 0.2% to 1.5%, with an increment of 0.1%). The fifth group consisted of 12 datasets with different coverage (200x, 300x, 400x and 500x) for each of the three species genomes. Since strains from the same reference genome may relate to each other with different evolutionary trajectories, we simulated another 162 datasets to mimic different evolutionary relationship in the sixth, seventh and eighth groups. We defined three types of similarity among strains: (1) Type1: For datasets with three strains, two strains are set to share 10% to 60% polymorphic sites, and the third one is independent to these two strains. The sixth group contained 54 datasets with Type1 evolutionary relationship. (2) Type2: strain1 and strain2 shared 10% to 60% of polymorphic sites, strain3 shared 10% sites with strain1 and strain2, while strain4 did not share polymorphic sites with the other strains. The seventh group contained 54 datasets with Type2 evolutionary relationship (3) Type3: strain1 and strain2 shared 10% to 60% of polymorphic sites, strain3 and strain4 shared 10% to 60% of polymorphic sites, while strain1 and strain2 shared no polymorphic sites with strain3 and strain4. The eighth group contained 54 datasets with Type3 evolutionary relationship.

The parameters and command to generate each simulated dataset were provided in the mixtureS tool package at http://www.cs.ucf.edu/~xiaoman/mixtureS/simulated243.

4.2.2 Experimental datasets

We also tested the tools on 195 experimental datasets. Sobkowiak et al. generated around 2000 datasets in their study and pointed out that 196 of these datasets were mixed with two strains (Sobkowiak et al., 2018). They inferred the abundance of the two strains with two different approaches based on a small number of known polymorphisms in these strains. Note that the polymorphic sites in strains were not available from this study. We thus only knew there were two strains and the strain proportion in each dataset. We managed to obtain 195 of the 196 datasets from the European Nucleotide Archive (Project ID ERP000436 and ERP001072). The 195 datasets were used to test whether the tools can identify two strains and how close the predicted strain abundance to the real strain abundance. The sample ID numbers of these 195 datasets were provided in the mixtureS tool at http://www.cs.ucf.edu/~xiaoman/mixtureS/experimental195.

4.2.3 MixtureS

As previous studies (Xin Li et al., 2019; Pulido-Tamayo et al., 2015; Smillie et al., 2018), mixtureS assumes that different strains of a species are likely to have different abundance. It also assumes that there are two types of nucleotides at a true polymorphic site, because almost all polymorphic sites in microbial genomes are biallelic (Foster, Bull, & Keim, 2020). The first assumption makes the separation of the polymorphic sites in different strains possible, and the second one enables a simpler solution as shown below. Note that although mixtureS needs the known species reference genome, it de novo infers the strain genomes and their variations and thus does not rely on known strains of a species or the known variations in strains. Moreover, the different strains are still allowed to share polymorphisms even with the second assumption. In addition, although there may

be more than two types of nucleotides at polymorphic positions, the exclusion of which is unlikely to affect the estimation of the strain number and the separation of the biallelic polymorphic sites from these unknown strains.

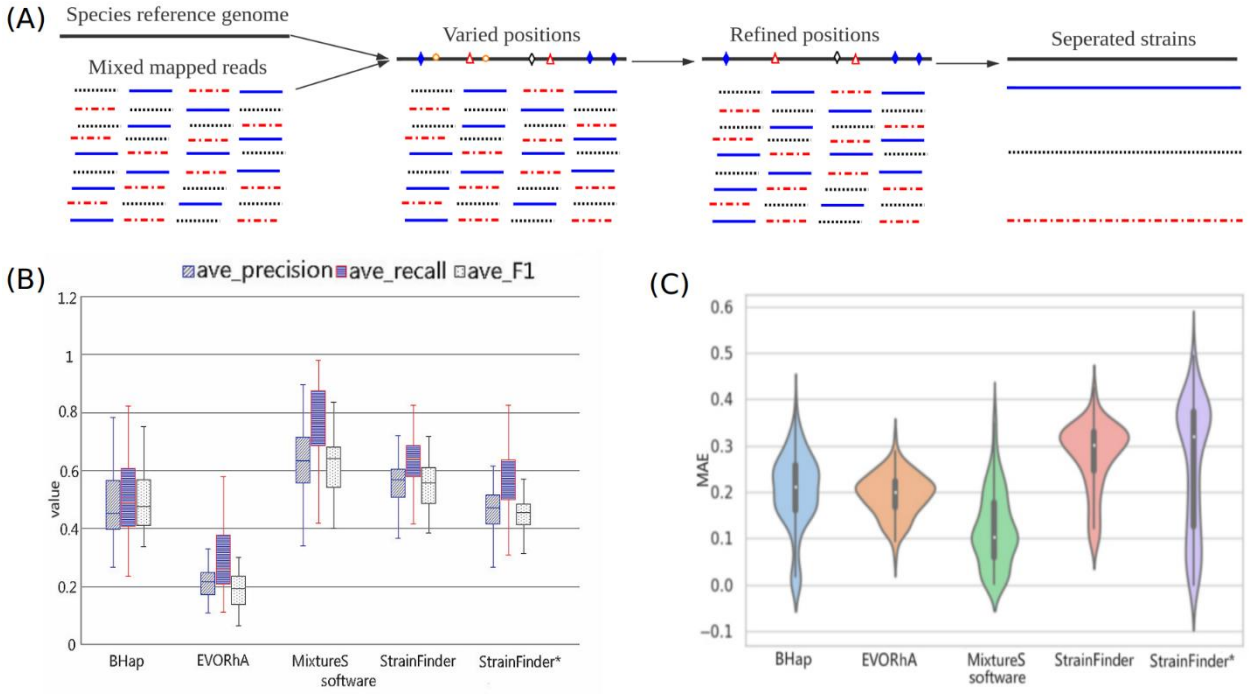

Figure 7 The mixtureS tool and its performance

(A) The three main steps in mixtureS. (B) Performance of mixtureS and other tools on simulated data. (C) Performance of mixtureS andother tools on experimental data. MAE on the y-axis is the average absolute difference between the predicted abundance of a predicted strain and the corresponding known abundance of thecorresponding known strain across strains and samples.

Starting from the mapped reads to a species genome, mixtureS infers the strain genomes and their abundance in three main steps (Figure 7A). First, it identifies all genome positions with varied nucleotides in the mapped reads. All genomic positions with two or more different nucleotides in the mapped reads are kept. Second, it refines the identified positions by removing positions with

low-coverage (< 10% of the average coverage of the genome) and positions with variations highly likely due to sequencing errors. The removal of low-coverage positions makes the remaining positions have more even coverage and also avoids the inference based on limited coverage. To remove positions with variations highly likely only because of sequencing errors, mixtureS chooses the two nucleotides with the two largest frequency at each of the remaining positions, in which the reference nucleotide should be one of the two nucleotides chosen. If the reference nucleotide is not chosen at a position, the nucleotide with the largest frequency in reads will be used to replace the reference nucleotide at this position. In this way, we have an n' by 2 matrix for the n' chosen potential polymorphic positions. We artificially normalize each row of this matrix so that the sum of the two frequencies in each row to be 100. In other words, we artificially create a uniform coverage of 100x across the n' positions. All positions with the smaller of the two frequencies smaller than 5 are removed, which correspond to the corrected p-value cutoff of 0.01 by assuming the sequencing error rate is 0.001 and the genome size is 1 Mbps. Assume we now have n positions left, which are from m strains. These positions are likely to contain the vast majority of true polymorphic sites, although there may still a small fraction of false polymorphic sites due to sequencing errors. MixtureS then sort the n rows of the remaining matrix according to the numbers in the second column from the smallest to the largest. Intuitively, the positions from the same strains are likely to correspond to rows next to each other in this sorted matrix. Finally, mixtureS applies an expectation maximization (EM) algorithm to infer the strains from the remaining polymorphic positions. EM algorithms have shown good performance previously (Xiaoman Li & Waterman, 2003; Smillie et al., 2018; Ying Wang et al., 2015, 2016, 2017).

In brief, assume that there are n remaining polymorphic positions, which are from m strains, and the frequency of the wide type nucleotide and the mutated nucleotide at the i-th position are $x_1^{(i)}$

and $x_2^{(i)}$, respectively. Assume the relative abundance of j-th strain is π_j and the probability that a mutated nucleotide at a position belongs to the j-th strain is α_j .We have the expectation of the missing data $\omega_j^{(i)}$at the E-step calculated as P_r (Function 7), where the missing data $z^{(i)}$ = j means the mutated nucleotide at the i-th position is from the j-th strain. We also define the Binomial probability function 8.

For the initialization of the EM algorithm, we set m=2 and then use the Jenks Natural Breaks to find the best place to divide the rows into two groups based on the 2nd column of the matrix, i.e., $\{x_2^{(i)}\}$. π_j is defined as normalized median with the j-th group and α_j is normalized as the number of rows in the j-th group, for j=1 or 2. We have the parameter estimation α_j and π_j at the M-step (Function 9 and Function 10). The E-step and M-step are iterated until the difference of the estimated parameters between two adjacent iterations is smaller than 1e-6 or the iteration number becomes 500.

In the above analysis, we start the EM algorithm with m=2. We compare the model with the model for m=1, whose calculation is straightforward, based on Bayesian information criterion (BIC). The model with the lower BIC is chosen. If the model from m=1 is chosen, we stop and report that there is only one strain. Otherwise, if the model from m=2 is chosen, we increase m by 1 to run the EM algorithm again. With the convergence of the EM algorithm at m=3, we obtain two models with different numbers of strains. We will then compare the two models based on BIC. The model with the lower BIC is chosen. If the model from m=2 is chosen, we will stop and report the two strains predicted with m=2. Otherwise, if the model from m=3 is chosen, this process is repeated until the better model is identified. In this way, we determine m, the strain number, together with

the strain abundance. The mutated nucleotide at the i-th position is assigned to the strain with the largest $w_j^{(i)}$based on the final model we chose.

$$P_r\left(\mathrm{z^{(i)}=j}\,\middle|\,\mathrm{x_1^{(i)}+x_2^{(i)}},\pi,\alpha\right)=\frac{B(\mathrm{x_1^{(i)}+x_2^{(i)}},\pi_j;\mathrm{x_2^{(i)}})*\alpha_j}{\sum_{k=1}^{n}B(\mathrm{x_1^{(i)}+x_2^{(i)}},\pi_k;\mathrm{x_2^{(i)}})*\alpha_k} \quad (7)$$

$$B\left(\mathrm{x_1^{(i)}+x_2^{(i)}},\pi_k;\mathrm{x_2^{(i)}}\right)=\frac{(\mathrm{x_1^{(i)}+x_2^{(i)}})!}{\mathrm{x_1^{(i)}!x_2^{(i)}!}}*(1-\pi_k)^{\mathrm{x_1^{(i)}}}(\pi_k)^{\mathrm{x_2^{(i)}}} \quad (8)$$

$$\alpha_j=\frac{1}{n}\sum_{i=1}^{n}w_j^{(i)} \quad (9)$$

$$\pi_j=\sum_{i=1}^{n}(x_2^{(i)}w_j^{(i)}/(x_1^{(i)}+x_2^{(i)}))/(\sum_{i=1}^{n}w_j^{(i)}) \quad (10)$$

We implement the above pipeline into the mixtureS tool package. This package provides the tool in both Linux and Windows versions. It also includes a script to process the FASTQ raw reads, to map reads to the reference genome of interest, and to generate a 4 by n matrix as the input to the above pipeline. The readme, source code, information about simulated and experimental datasets, together with example test datasets, are also provided. Compared with existing tools (Pulido-Tamayo et al., 2015; Smillie et al., 2018; Truong et al., 2017), it is much easier to set up the running environment for mixtureS and simpler to interpret the output of mixtureS, which makes it easy to apply mixtureS for strain studies.

4.2.4 Comparison with BHap, EOVRhA and strainFinder

We test mixtureS, BHap, EVORhA, and strainFinder on the above simulated and experimental datasets. Because mixtureS and EVORhA required the bam format input, we used trimmomatic to trim raw fastq data, and then used bowtie2 to map raw reads into sam format file. The bam format file were obtained from the sam format file by converting it with SamTools. The mixtureS was run by command: python mixture_model.py --sample_name --genome_len --genome_name genome_nc_name --genome_file_loc genome.fna --bam_file input_sorted.bam --res_dir ./res_dir.

BHap is also run by the default parameter and the command: python ./run_BHap.py -d ./res_name -r genome.fna -t fastq -1 forward.fastq -2 reverse.fastq -l read_len -c coverage -g genome_len. EVORhA was also run by the default parameter with java -Xmx120g -Xms40g -XX:-UseGCOverheadLimit -jar evorha.jar completeAnalysis genome_name bam_name. Because StrainFinder requires Numpy array as input, we exactly follow their pre-processing script in the StrainFinder tool set to generate Numpy array. StrainFinder was run with the default parameter that would try to predict ten strains, and the command is python StrainFinder.py --aln in_gene_aln -N 10 --max_reps 10 --dtol 1 –ntol 2 --max_time 3600 –converge –em res_em --em_out res_em --otu_out res_table –log res_log --n_keep 3 --force_update –msg --merge_out. StrainFinder* was used the same parameter but with parameter -N set to be the correct number of strains.

To compare the performance of the tools on simulated datasets, we considered the number of strains predicted and the absolute difference of between the predicted abundance and the true abundance of the strains. We called the absolute difference of between the predicted abundance and the true abundance of the strains MAE as previously (Xin Li et al., 2019; Pulido-Tamayo et al., 2015). We also considered the precision, recall and F1 score of the predicted polymorphic sites in strains. Note that when a tool predicted more than m strains, where m is the true number of strains in this dataset, only the m predicted strains that shared the largest number of polymorphic sites with the m known strains were used for comparison. When a tool predicted fewer than m strains, say k strains, only the predicted k strains that shared the largest number of polymorphic sites with k out of m known strains were used for comparison. This implies that an advantage gives to tools that predicted more strains. In this study, the order of the tools predicted more strains in a dataset is StrainFinder, EVORhA, BHap and MixtureS. We also compared the folds of improvement by mixtureS compared with other methods. To calculate the folds of improvement,

we obtained the MAE for each method in each sample. We then calculated the fold of improvement in a sample as the ratio of the MAE of a method to the MAE of mixtureS. When the MAE of mixtureS is 0 in a sample, which means that mixtureS perfectly predicted the correct abundance in this sample, this sample is not considered. Finally, we averaged the fold of improvement across sample for a method compared with mixtureS as the fold of improvement of mixtureS compared with this method.

To compare the performance of the tools on the 195 experimental datasets, we considered the number of strains predicted and the MAE of the predicted strains. Since we do not know the polymorphic sites in strains, we cannot calculate the precision, recall, or F1 scores. When a tool predicted more than 2 strains, only the 2 predicted strains that had most similar abundance with the abundance of the 2 known strains were used for comparison. When a tool predicted only one strain, only this predicted strain was compared with the known strain with more similar abundance to this predicted strain.

4.3 Results

4.3.1 mixtureS has best performance on simulated datasets

We tested mixtureS on 243 simulated datasets (Table 18, Supplementary S2A in original paper) (Xin Li, Hu, & Li, 2020). In each dataset, we randomly generated shotgun reads for 2 to 4 strains of a bacterial species and mixed these reads together. We tested mixtureS together with three other tools, BHap, EVORhA, and strainFinder on the mixed reads (Xin Li et al., 2019; Pulido-Tamayo et al., 2015; Smillie et al., 2018). MixtureS predicted the correct strain numbers in 202 datasets, while BHap, EVORhA, and strainFinder did it in 40, 46, and 0 datasets, respectively. Because strainFinder had trouble to find the right strain number, we input the correct strain number to run

strainFinder, which was called strainFinder*. Even with this advantage, in terms of the strain abundance, mixtureS predicted at least 2.96, 1.74, 7.68 and 3.71 times closer to the true abundance than BHap, EVORhA, strainFinder, and strainFinder*, respectively (the corresponding standard deviation as 8,4, 40.70 and 18.50, respectively). In addition, the predicted polymorphic sites by mixtureS was much more accurate (Figure 7B).

Table 18 Summary Results on simulated datasets

measurement	Bhap	EVORhA	StrainFinder	StrainFinder*	MixtureS
MAE without evolutionary	0.013 (0.007)	0.033 (0.036)	0.361 (0.131)	0.067 (0.061)	0.008 (0.001)
MAE with evolutionary	0.132 (0.089)	0.065 (0.057)	0.177 (0.094)	0.139 (0.073)	0.042 (0.044)
MAE overall	0.092 (0.094)	0.054 (0.06)	0.238 (0.138)	0.115 (0.077)	0.031 (0.041)
F1-score	0.513	0.226	0.634	0.525	0.718
Abs difference between # of real strains and # of predicted strains	1.239 (0.612)	1.617 (1.077)	3.770 (1.163)	0.008 (0)	0.280 (0)

MAE is the average absolute difference between the predicted abundance of the predicted strains and the true abundance of the most similar corresponding known strains. The numbers in the parentheses are the corresponding standard deviation.

4.3.2 mixtureS has best performance on experimental datasets

We also tested mixtureS on 195 experimental datasets (Table 19 and Table 20) (Sobkowiak et al., 2018). There were two strains of Mycobacterium tuberculosis with known abundance in each dataset, while the polymorphic sites in the two strains were unknown. We compared how well the four methods predicted the number of strains and their abundance. BHap, EVORhA, strainFinder, and mixtureS predicted two strains in 22, 0, 0 and 84 datasets, respectively. As to the strain abundance, mixtureS had a much accurate estimate than other tools, including strainFinder* (Figure 7C).

Table 19 Summary results on experimental datasets

measurements	BHap	EVORhA	StrainFinder	StrainFinder*	MixtureS
ave # of reconstructed strains	4.174 (1.533)	5.113 (0.918)	7.211 (1.367)	1.969 (0)	2.862 (0.895)
Average absolute difference of the abundance between the predicted strains and the corresponding known strains	0.202 (0.081)	0.194 (0.045)	0.279 (0.073)	0.259 (0.139)	0.121 (0.079)

The result is tested on 195 experimental datasets. The numbers in the parentheses are the corresponding standard deviation.

Table 20 Results on each experimental dataset

sample	Verified by any of reference paper method	major strain proportion	Bhap	EVORhA	StrainFinder	StrainFinder*	MixtureS
ERR036233	both	0.72	2(0.008)	7(0.186)	8(0.243)	2(0.215)	4(0.113)
ERR036248	both	0.88	2(0.09)	6(0.271)	9(0.331)	2(0.114)	2(0.01)
ERR037469	both	0.63	9(0.307)	5(0.073)	4(0.136)	2(0.115)	2(0.144)
ERR037547	both	0.85	10(0.335)	7(0.18)	1(0.15)	2(0.035)	2(0.042)
ERR126641	both	0.84	5(0.224)	7(0.224)	8(0.216)	2(0.063)	4(0.137)
ERR126642	both	0.8	5(0.202)	8(0.189)	10(0.326)	2(0.055)	2(0.088)
ERR161024	both	0.86	3(0.158)	5(0.188)	8(0.328)	2(0.003)	3(0.082)
ERR161026	both	0.85	3(0.17)	4(0.099)	9(0.312)	2(0.066)	3(0.106)
ERR161027	both	0.82	4(0.168)	5(0.205)	8(0.273)	2(0.046)	3(0.06)
ERR161034	both	0.65	4(0.097)	5(0.119)	5(0.236)	2(0.067)	3(0.086)
ERR161039	both	0.63	4(0.119)	5(0.161)	8(0.256)	2(0.084)	4(0.157)
ERR161049	both	0.87	2(0.059)	5(0.191)	8(0.125)	2(0.246)	3(0.099)
ERR161050	both	0.73	2(0.002)	5(0.155)	9(0.308)	2(0.214)	5(0.218)
ERR161055	both	0.89	4(0.225)	6(0.27)	5(0.214)	2(0.387)	2(0.005)
ERR161071	both	0.84	4(0.23)	6(0.207)	1(0.16)	2(0.008)	4(0.148)
ERR161077	both	0.78	3(0.185)	5(0.143)	9(0.286)	2(0.206)	5(0.238)
ERR161078	both	0.58	4(0.162)	5(0.118)	6(0.263)	2(0.038)	4(0.123)
ERR161081	both	0.88	5(0.257)	4(0.144)	6(0.324)	2(0.317)	2(0.124)
ERR161084	both	0.87	3(0.163)	7(0.256)	5(0.293)	2(0.047)	3(0.103)
ERR161088	both	0.88	2(0.079)	5(0.231)	8(0.349)	1(0.12)	3(0.11)
ERR161090	both	0.91	4(0.214)	4(0.211)	7(0.331)	2(0.406)	2(0.039)
ERR161091	both	0.89	4(0.21)	5(0.205)	6(0.296)	2(0.336)	2(0.003)
ERR161097	both	0.85	5(0.261)	7(0.229)	8(0.307)	2(0.057)	2(0.025)
ERR161120	both	0.86	5(0.261)	5(0.221)	6(0.301)	2(0.177)	2(0.104)
ERR161122	both	0.87	2(0.002)	5(0.237)	6(0.086)	2(0.239)	3(0.099)
ERR161123	both	0.81	3(0.181)	4(0.17)	7(0.305)	2(0.042)	3(0.056)
ERR161170	both	0.89	3(0.153)	4(0.203)	4(0.22)	2(0.352)	2(0.228)
ERR161173	both	0.9	3(0.23)	4(0.248)	5(0.242)	2(0.378)	2(0.265)
ERR161176	both	0.88	3(0.184)	4(0.205)	9(0.332)	2(0.375)	3(0.1)

sample	Verified by any of reference paper method	major strain proportion	Bhap	EVORhA	StrainFinder	StrainFinder*	MixtureS
ERR161184	both	0.87	3(0.156)	5(0.184)	5(0.252)	2(0.349)	3(0.196)
ERR161194	both	0.86	2(0.0)	6(0.247)	7(0.091)	2(0.121)	3(0.101)
ERR161195	both	0.88	5(0.262)	4(0.137)	8(0.169)	2(0.377)	2(0.01)
ERR181749	both	0.87	3(0.154)	4(0.21)	3(0.196)	2(0.285)	2(0.209)
ERR181750	both	0.89	3(0.157)	4(0.214)	8(0.298)	2(0.359)	2(0.078)
ERR181752	both	0.85	4(0.197)	5(0.23)	9(0.313)	2(0.317)	2(0.007)
ERR181753	both	0.87	3(0.166)	4(0.182)	9(0.318)	1(0.13)	2(0.004)
ERR181782	both	0.86	4(0.262)	6(0.192)	5(0.193)	2(0.042)	3(0.085)
ERR181784	both	0.82	5(0.234)	6(0.203)	6(0.239)	2(0.047)	3(0.055)
ERR181785	both	0.8	6(0.243)	7(0.227)	7(0.276)	2(0.027)	2(0.086)
ERR181810	both	0.9	3(0.161)	5(0.186)	8(0.342)	2(0.375)	2(0.23)
ERR181811	both	0.54	2(0.17)	4(0.125)	8(0.263)	2(0.0)	5(0.191)
ERR181813	both	0.62	4(0.136)	5(0.153)	10(0.275)	2(0.084)	4(0.121)
ERR181827	both	0.88	3(0.17)	4(0.126)	8(0.339)	2(0.374)	3(0.096)
ERR181828	both	0.88	4(0.235)	5(0.175)	7(0.342)	2(0.319)	3(0.092)
ERR181838	both	0.88	3(0.176)	6(0.275)	9(0.344)	2(0.378)	2(0.004)
ERR181845	both	0.9	5(0.236)	4(0.205)	7(0.327)	2(0.397)	2(0.035)
ERR181849	both	0.88	3(0.156)	5(0.195)	9(0.34)	2(0.361)	3(0.095)
ERR181866	both	0.89	3(0.159)	4(0.217)	7(0.325)	1(0.11)	3(0.104)
ERR181870	both	0.9	4(0.229)	5(0.127)	6(0.253)	2(0.398)	2(0.024)
ERR181876	both	0.91	5(0.268)	4(0.172)	2(0.201)		
ERR181878	both	0.81	3(0.135)	5(0.21)	6(0.254)	2(0.293)	2(0.101)
ERR181880	both	0.86	3(0.17)	4(0.201)	5(0.182)	2(0.341)	2(0.011)
ERR181881	both	0.84	2(0.019)	5(0.186)	7(0.305)	2(0.129)	4(0.146)
ERR181909	both	0.89	4(0.26)	5(0.217)	7(0.333)	2(0.378)	2(0.037)
ERR181913	both	0.89	3(0.152)	5(0.228)	7(0.32)	2(0.362)	2(0.179)
ERR181923	both	0.88	4(0.191)	4(0.177)	9(0.332)	1(0.12)	2(0.015)
ERR181933	both	0.91	4(0.228)	4(0.169)	7(0.334)	2(0.409)	2(0.043)
ERR181937	both	0.87	4(0.248)	5(0.246)	8(0.15)	2(0.349)	2(0.012)
ERR176620	both	0.54	3(0.103)	4(0.08)	8(0.259)	2(0.015)	3(0.096)
ERR176621	both	0.9	5(0.246)	5(0.246)	8(0.303)	2(0.273)	2(0.002)
ERR176631	both	0.89	3(0.171)	5(0.216)	10(0.358)	2(0.388)	3(0.099)
ERR176650	both	0.87	4(0.264)	6(0.202)	8(0.334)	2(0.363)	4(0.152)
ERR176652	both	0.82	2(0.03)	6(0.231)	9(0.306)	2(0.304)	5(0.245)
ERR176653	both	0.8	2(0.004)	5(0.199)	7(0.29)	2(0.29)	4(0.178)
ERR176655	both	0.91	4(0.251)	4(0.119)	6(0.323)	2(0.403)	2(0.208)
ERR176664	both	0.88	3(0.152)	4(0.184)	6(0.265)	2(0.378)	3(0.229)
ERR176668	both	0.92	6(0.308)	5(0.247)	8(0.345)	2(0.395)	2(0.276)
ERR176672	both	0.89	4(0.209)	4(0.205)	7(0.289)	2(0.371)	2(0.237)
ERR176681	both	0.89	5(0.22)	4(0.108)	7(0.32)	2(0.388)	2(0.187)
ERR176688	both	0.9	3(0.157)	4(0.178)	6(0.3)	2(0.381)	2(0.253)
ERR176701	both	0.83	4(0.218)	4(0.196)	8(0.148)	2(0.327)	4(0.165)
ERR176706	both	0.89	5(0.237)	4(0.155)	7(0.322)	2(0.362)	2(0.178)
ERR176709	both	0.65	6(0.253)	5(0.178)	9(0.282)	2(0.142)	5(0.171)

sample	Verified by any of reference paper method	major strain proportion	Bhap	EVORhA	StrainFinder	StrainFinder*	MixtureS
ERR176713	both	0.89	2(0.121)	4(0.143)	8(0.321)	2(0.345)	4(0.194)
ERR176723	both	0.88	4(0.251)	6(0.246)	7(0.312)	2(0.361)	3(0.103)
ERR176725	both	0.88	2(0.023)	5(0.238)	9(0.335)	2(0.37)	4(0.152)
ERR176734	both	0.91	4(0.245)	5(0.222)	9(0.336)	2(0.375)	2(0.017)
ERR176738	both	0.88	3(0.151)	5(0.246)	6(0.26)	2(0.379)	4(0.193)
ERR176746	both	0.87	4(0.242)	4(0.148)	8(0.328)	2(0.359)	2(0.176)
ERR176748	both	0.87	4(0.188)	4(0.151)	9(0.345)	2(0.09)	2(0.031)
ERR176749	both	0.81	5(0.223)	5(0.113)	9(0.299)	2(0.023)	2(0.035)
ERR176755	both	0.89	4(0.239)	3(0.095)	7(0.337)	2(0.375)	2(0.179)
ERR176793	both	0.54	4(0.11)	4(0.117)	9(0.275)	2(0.004)	5(0.198)
ERR176796	both	0.88	4(0.272)	5(0.214)	7(0.316)	2(0.374)	2(0.118)
ERR176802	both	0.89	4(0.239)	4(0.21)	7(0.337)	2(0.347)	2(0.217)
ERR176807	both	0.87	2(0.009)	5(0.198)	6(0.104)	2(0.367)	3(0.098)
ERR176809	both	0.89	3(0.152)	5(0.191)	6(0.292)	2(0.387)	2(0.202)
ERR176810	both	0.91	4(0.229)	4(0.172)	9(0.347)	2(0.409)	2(0.214)
ERR176813	both	0.88	4(0.246)	5(0.214)	7(0.341)	2(0.377)	3(0.106)
ERR181686	both	0.87	3(0.179)	5(0.221)	6(0.198)	2(0.079)	3(0.094)
ERR181688	both	0.83	3(0.134)	5(0.126)	7(0.107)	2(0.249)	2(0.041)
ERR181689	both	0.85	4(0.186)	4(0.12)	8(0.323)	2(0.069)	2(0.044)
ERR181695	both	0.87	3(0.162)	6(0.245)	7(0.322)	2(0.368)	2(0.013)
ERR181705	both	0.84	5(0.251)	5(0.204)	9(0.34)	2(0.028)	2(0.044)
ERR216914	both	0.75	6(0.237)	5(0.179)	8(0.3)	2(0.225)	5(0.235)
ERR216917	both	0.89	7(0.284)	6(0.215)	7(0.328)	2(0.381)	4(0.201)
ERR216932	both	0.89	5(0.23)	6(0.213)	5(0.198)	2(0.387)	3(0.233)
ERR216933	both	0.9	6(0.29)	4(0.114)	7(0.243)	2(0.341)	2(0.243)
ERR216942	both	0.87	7(0.315)	5(0.191)	9(0.338)	2(0.366)	2(0.019)
ERR216952	both	0.91	7(0.311)	5(0.233)	7(0.345)	1(0.09)	4(0.228)
ERR216956	both	0.89	5(0.268)	6(0.232)	10(0.322)	2(0.386)	3(0.153)
ERR216961	both	0.93	7(0.299)	6(0.227)	7(0.21)	2(0.392)	3(0.217)
ERR216966	both	0.88	6(0.238)	7(0.253)	9(0.337)	2(0.316)	3(0.131)
ERR216967	both	0.88	7(0.268)	6(0.194)	6(0.324)	2(0.376)	3(0.13)
ERR216971	both	0.69	6(0.215)	6(0.146)	8(0.293)	2(0.187)	6(0.211)
ERR216974	both	0.89	6(0.27)	6(0.212)	6(0.091)	2(0.223)	3(0.145)
ERR216977	both	0.89	8(0.298)	5(0.137)	7(0.119)	2(0.387)	3(0.167)
ERR216983	both	0.89	5(0.224)	7(0.218)	9(0.333)	2(0.386)	3(0.139)
ERR216984	both	0.88	7(0.262)	6(0.245)	6(0.239)	2(0.374)	3(0.123)
ERR216989	both	0.87	7(0.269)	6(0.178)	5(0.278)	2(0.238)	3(0.151)
ERR221524	both	0.88	4(0.233)	4(0.119)	9(0.166)	2(0.336)	2(0.005)
ERR221536	both	0.87	2(0.018)	6(0.258)	9(0.349)	2(0.249)	3(0.101)
ERR221538	both	0.88	5(0.269)	6(0.222)	7(0.232)	2(0.368)	4(0.16)
ERR221539	both	0.82	4(0.188)	5(0.17)	8(0.329)	2(0.105)	3(0.073)
ERR221561	both	0.69	5(0.189)	6(0.165)	9(0.318)	2(0.189)	4(0.115)
ERR221567	both	0.87	6(0.308)	6(0.271)	9(0.345)	2(0.369)	3(0.107)
ERR221611	both	0.87	3(0.157)	4(0.135)	7(0.307)	2(0.357)	3(0.102)

sample	Verified by any of reference paper method	major strain proportion	Bhap	EVORhA	StrainFinder	StrainFinder*	MixtureS
ERR245754	both	0.57	10(0.342)	8(0.156)	5(0.114)	2(0.007)	3(0.093)
ERR245758	both	0.79	9(0.317)	6(0.169)	1(0.21)	2(0.038)	2(0.078)
ERR245795	both	0.65	8(0.315)	6(0.129)	9(0.305)	2(0.147)	2(0.112)
ERR245797	both	0.82	7(0.282)	6(0.19)	6(0.289)	2(0.054)	2(0.07)
ERR323044	both	0.71	3(0.045)	4(0.141)	8(0.273)	2(0.208)	2(0.052)
ERR323054	both	0.66	5(0.198)	4(0.105)	6(0.179)	2(0.152)	5(0.192)
ERR323082	both	0.71	4(0.144)	6(0.169)	8(0.28)	2(0.162)	4(0.114)
ERR473322	both	0.77	2(0.012)	3(0.051)	8(0.292)	2(0.237)	4(0.195)
ERR473359	both	0.5	2(0.156)	6(0.221)	7(0.261)	2(0.017)	3(0.063)
ERR773806	both	0.91	2(0.258)	4(0.241)	8(0.29)	2(0.353)	4(0.3)
ERR181953	both	0.87	5(0.261)	6(0.214)	8(0.305)	2(0.368)	2(0.013)
ERR181974	both	0.85	4(0.196)	6(0.257)	6(0.09)	2(0.219)	3(0.094)
ERR181977	both	0.8	4(0.206)	7(0.257)	8(0.283)	2(0.056)	4(0.143)
ERR181983	both	0.9	4(0.228)	5(0.26)	6(0.334)	2(0.373)	3(0.094)
ERR182015	both	0.85	4(0.17)	5(0.169)	8(0.325)	2(0.176)	4(0.143)
ERR182026	both	0.84	4(0.234)	4(0.135)	7(0.319)	2(0.064)	4(0.151)
ERR182027	both	0.87	3(0.163)	5(0.173)	7(0.32)	2(0.167)	4(0.138)
ERR182041	both	0.88	5(0.251)	4(0.145)	8(0.148)	2(0.368)	2(0.011)
ERR182049	both	0.89	3(0.15)	4(0.216)	6(0.34)	2(0.366)	3(0.098)
ERR190340	both	0.63	1(0.37)	5(0.183)	4(0.143)	2(0.127)	5(0.225)
ERR190342	both	0.86	3(0.177)	6(0.179)	7(0.301)	2(0.165)	3(0.09)
ERR190343	both	0.8	5(0.21)	6(0.231)	8(0.284)	2(0.029)	3(0.045)
ERR190379	both	0.77	3(0.191)	6(0.175)	7(0.286)	2(0.2)	4(0.111)
ERR190388	both	0.91	4(0.211)	4(0.255)	5(0.261)	1(0.09)	3(0.268)
ERR211990	both	0.89	3(0.148)	5(0.218)	7(0.314)	2(0.343)	3(0.101)
ERR212002	both	0.86	2(0.003)	4(0.165)	8(0.143)	2(0.323)	3(0.106)
ERR212004	both	0.86	2(0.002)	4(0.219)	6(0.314)	2(0.336)	4(0.157)
ERR212041	both	0.85	6(0.234)	5(0.183)	5(0.301)	2(0.337)	2(0.162)
ERR212058	both	0.88	3(0.176)	4(0.219)	8(0.316)	2(0.349)	3(0.11)
ERR212059	both	0.86	3(0.17)	5(0.212)	8(0.314)	2(0.356)	3(0.11)
ERR212069	both	0.86	3(0.173)	5(0.212)	8(0.285)	2(0.312)	3(0.092)
ERR212086	both	0.88	7(0.334)	6(0.199)	6(0.262)	2(0.129)	2(0.021)
ERR212098	both	0.84	6(0.265)	7(0.221)	8(0.309)	2(0.066)	3(0.073)
ERR212100	both	0.84	5(0.255)	6(0.204)	8(0.338)	2(0.027)	3(0.073)
ERR212101	both	0.85	9(0.325)	7(0.205)	8(0.243)	2(0.009)	3(0.076)
ERR212107	both	0.87	4(0.276)	5(0.202)	6(0.222)	2(0.211)	3(0.092)
ERR212112	both	0.89	3(0.183)	5(0.24)	9(0.319)	2(0.357)	4(0.191)
ERR212134	both	0.85	6(0.261)	5(0.137)	6(0.28)	2(0.328)	2(0.007)
ERR212161	both	0.87	5(0.26)	5(0.169)	6(0.102)	2(0.336)	3(0.102)
ERR212165	both	0.85	6(0.28)	4(0.177)	7(0.121)	2(0.266)	3(0.09)
ERR216899	both	0.88	7(0.312)	6(0.211)	8(0.314)	2(0.367)	4(0.212)
ERR163932	both	0.91	3(0.154)	7(0.289)	7(0.349)	2(0.401)	3(0.097)
ERR176616	both	0.63	3(0.1)	7(0.208)	9(0.293)	2(0.125)	5(0.191)
ERR181708	both	0.87	3(0.174)	6(0.217)	10(0.33)	2(0.321)	3(0.092)

sample	Verified by any of reference paper method	major strain proportion	Bhap	EVORhA	StrainFinder	StrainFinder*	MixtureS
ERR181945	both	0.87	4(0.188)	6(0.193)	7(0.306)	2(0.364)	4(0.145)
ERR216913	both	0.88	6(0.254)	5(0.155)	7(0.317)	2(0.371)	2(0.004)
ERR163940	both	0.87	4(0.207)	5(0.171)	6(0.286)	2(0.364)	2(0.04)
ERR163942	both	0.87	3(0.159)	6(0.218)	8(0.329)	2(0.144)	3(0.094)
ERR163943	both	0.83	4(0.196)	5(0.196)	7(0.106)	2(0.101)	3(0.087)
ERR163947	both	0.5	3(0.136)	5(0.188)	10(0.291)	2(0.04)	3(0.122)
ERR163954	both	0.92	4(0.22)	4(0.194)	8(0.355)	2(0.417)	2(0.155)
ERR163971	both	0.89	4(0.227)	5(0.246)	8(0.332)	2(0.388)	2(0.007)
ERR163986	both	0.9	4(0.211)	4(0.161)	7(0.351)	2(0.394)	2(0.018)
ERR163996	both	0.88	4(0.2)	6(0.255)	7(0.329)	2(0.377)	3(0.098)
ERR164007	both	0.88	3(0.164)	5(0.18)	7(0.104)	2(0.167)	2(0.01)
ERR164021	both	0.7	6(0.212)	6(0.155)	7(0.23)	2(0.187)	3(0.043)
ERR176446	both	0.88	6(0.294)	7(0.247)	7(0.328)	2(0.379)	4(0.152)
ERR176458	both	0.82	3(0.163)	6(0.223)	8(0.294)	2(0.316)	4(0.16)
ERR176460	both	0.88	3(0.174)	5(0.219)	7(0.13)	2(0.373)	2(0.001)
ERR176461	both	0.8	4(0.199)	5(0.175)	7(0.273)	2(0.298)	2(0.06)
ERR176521	both	0.89	5(0.257)	5(0.244)	7(0.339)	2(0.385)	2(0.132)
ERR176533	both	0.9	5(0.23)	5(0.144)	7(0.329)	2(0.275)	2(0.001)
ERR176549	both	0.72	5(0.179)	5(0.113)	5(0.219)	2(0.213)	2(0.053)
ERR176556	both	0.86	5(0.267)	5(0.154)	7(0.214)	2(0.113)	2(0.041)
ERR176557	both	0.86	5(0.254)	6(0.198)	9(0.335)	2(0.013)	2(0.058)
ERR176600	both	0.89	5(0.273)	6(0.216)	6(0.297)	2(0.386)	2(0.014)
ERR176604	both	0.89	3(0.187)	5(0.245)	9(0.326)	2(0.369)	3(0.111)
ERR176610	both	0.9	4(0.246)	4(0.154)	8(0.335)	2(0.397)	2(0.026)
ERR176611	both	0.88	2(0.024)	4(0.12)	9(0.176)	2(0.376)	3(0.1)
ERR036194	single	1	1(0.0)	4(0.24)	9(0.412)	2(0.459)	2(0.349)
ERR176703	single	1	4(0.294)	5(0.326)	7(0.403)	2(0.493)	2(0.348)
ERR176785	single	1	6(0.353)	5(0.235)	6(0.406)	2(0.479)	2(0.32)
ERR221534	single	1	5(0.338)	5(0.277)	8(0.384)	2(0.465)	2(0.102)
ERR221592	single	1	4(0.296)	5(0.214)	7(0.2)	2(0.478)	2(0.25)
ERR245716	single	1	6(0.398)	6(0.307)	6(0.399)	2(0.29)	2(0.279)
ERR323056	single	1	4(0.317)	4(0.21)	9(0.364)	2(0.484)	4(0.348)
ERR473340	single	1	2(0.303)	5(0.277)	6(0.364)	2(0.437)	5(0.38)
ERR176514	single	1	4(0.258)	5(0.245)	4(0.423)	2(0.48)	2(0.282)

In each method column, the information in order is # of strains reconstructed, (average absolute abundance difference between the predicted strains and the known corresponding strains).

4.4 Discussion

We demonstrated the usage of mixtureS on samples of individual species. For metagenomic samples with multiple species, users can map reads to the species genome of interest first and then apply mixtureS. MixtureS can infer strains more accurately than existing tools and is fast (Table 21), which makes it a valuable addition to study bacterial strains.

Table 21 Running time comparison

software	Default dataset1 running time(seconds)	Default dataset1 running time(seconds)	Default dataset1 running time(seconds)	Average default dataset running time(seconds)	Average default dataset running time per 1 billion 100bp long reads(Hours)
BHap	136	108	248	164	10.0
EVORhA	1103	800	1869	1257	77.1
StrainFinder	517	582	831	643	41.8
StrainFinder*	84	80	252	139	7.9
MixtureS	266	221	505	331	20.2
StrainFinder preprocessing	613	500	1117	743	45.6
MixtureS preprocessing	453	350	777	527	32.4
EVORhA and Preprocessing	1556	1149	2646	1784	109.6
StrainFinder and Preprocessing	1129	1082	1948	1386	87.4
StrainFinder* and Preprocessing	696	580	1369	882	53.5
MixtureS and Preprocessing	719	571	1283	858	52.7

CHAPTER FIVE: CONCLUSIONS

5.1 Conclusion

Bacterial strain reconstruction is one of critical steps in metagenomics research. By understanding diversity of bacterial strain, we can have a better understanding of our earth, and select proper treatment for disease caused by corresponding strains. Reconstructing bacterial strain is a key step to identify the strain. However, existing methods are mostly working on viral, depending on known strains or not easy implemented.

In this dissertation, we study the affection of the newly sequenced genome and then present two approaches that can reconstruct bacterial strain. In Chapter 2, we found the inconsistent result from shotgun and 16s rRNA sequencing. Another inconsistent result can also be shown by using updated databases and tools in metagenomics. In this study, we also identify more potentially colitis related taxa by reanalyzing sequence. Both inconsistent results shed light on limitations of current genome databases and popular tools. In Chapter 3, we present a novel bacterial strain reconstruction method by fuzzy flow networks and the De Bruijn graph. BHap decomposed the fuzzy flow network and found feasible flows as strains. It shows robust performance under different parameters. In Chapter 4, another tool mixtureS was developed for reconstructing bacterial strain. Unlike BHap that is based on the De Bruijn graph to extend reads and decompose the strains by fuzzy flow networks, mixtureS directly mapped all reads into reference genome. Then mixtureS redefined all identified positions by filtering positions with low-coverage and sequence errors. mixtureS will apply an expectation maximization(EM) only on two nucleotides with two largest frequencies. At last, Bayesian information criterion(BIC) was used to estimate the correct number of strains.

5.2 Future work

For Chapter 2, we have found that some multi-reads may affect the analysis result. It may require tools that can better assign multi-reads to its real destination genome. For those taxa with nonsignificant on unique reads only but significant on all mapped reads, we do not have high confidence about their colitis-relatedness. Some of them may be colitis related because all its lower taxa are statistically significant related to colitis.

For BHap in Chapter 3, although BHap has better performance than EVORhA, some polymorphic sites in experimental datasets can not be defined correctly. It may still have a room for such improvement. Since BHap is based on coverage difference to distinguish strains, although it has robust performance under different parameters if there is different coverage between strains, the F1 score is much slower if two strains have the same coverage, such 50/50. That may be another aspect for improvement of BHap.

REFERENCES

Ahn, T.-H., Chai, J., & Pan, C. (2015). Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics, 31*(2), 170-177.

Albanese, D., & Donati, C. (2017). Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nature communications, 8*(1), 1-14.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., . . . Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods, 11*(11), 1144-1146.

Amann, R. I., Ludwig, W., & Schleifer, K.-H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews, 59*(1), 143-169.

Astrovskaya, I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., & Zelikovsky, A. (2011). *Inferring viral quasispecies spectra from 454 pyrosequencing reads.* Paper presented at the BMC bioinformatics.

Azuma, Y., Hirakawa, H., Yamashita, A., Cai, Y., Rahman, M. A., Suzuki, H., . . . Murakami, T. (2006). Genome sequence of the cat pathogen, Chlamydophila felis. *DNA research, 13*(1), 15-23.

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., & Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science, 307*(5717), 1915-1920.

Barrick, J. E., & Lenski, R. E. (2009). *Genome-wide mutational diversity in an evolving population of Escherichia coli.* Paper presented at the Cold Spring Harbor symposia on quantitative biology.

Bartlett, J. G., Onderdonk, A. B., Cisneros, R. L., & Kasper, D. L. (1977). Clindamycin-associated colitis due to a toxin-producing species of Clostridium in hamsters. *Journal of Infectious Diseases, 136*(5), 701-705.

Bäumler, A. J., & Sperandio, V. (2016). Interactions between the microbiota and pathogenic bacteria in the gut. *Nature, 535*(7610), 85-93.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr Opin Genet Dev, 16*(6), 545-552.

Berg, R. D. (1996). The indigenous gastrointestinal microflora. *Trends in microbiology, 4*(11), 430-435.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., . . . Ussery, D. W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & integrative genomics, 6*(3), 165-185.

Bloom, S. M., Bijanki, V. N., Nava, G. M., Sun, L., Malvin, N. P., Donermeyer, D. L., . . . Stappenbeck, T. S. (2011). Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell host & microbe, 9*(5), 390-403.

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods, 6*(9), 673-676.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., . . . Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences, 99*(22), 14250-14255.

Brenner, D. J., Staley, J. T., & Krieg, N. R. (2005). Classification of procaryotic organisms and the concept of bacterial speciation. In *Bergey's Manual «of Systematic Bacteriology* (pp. 27-32): Springer.

Bryant, J., Chewapreecha, C., & Bentley, S. D. (2012). Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future microbiology, 7*(11), 1283-1296.

Campieri, M., & Gionchetti, P. (2001). Bacteria as the cause of ulcerative colitis. *Gut, 48*(1), 132-135.

Chang, C., & Lin, H. (2016). Dysbiosis in gastrointestinal disorders. *Best practice & research Clinical gastroenterology, 30*(1), 3-15.

Chatterji, S., Yamazaki, I., Bai, Z., & Eisen, J. A. (2008). *CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads.* Paper presented at the Annual International Conference on Research in Computational Molecular Biology.

Collins, F. S., & McKusick, V. A. (2001). Implications of the Human Genome Project for medical science. *Jama, 285*(5), 540-544.

Connon, S. A., & Giovannoni, S. J. (2002). High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. *Applied and environmental microbiology, 68*(8), 3878-3885.

Darmon, E., & Leach, D. R. (2014). Bacterial genome instability. *Microbiology and molecular biology reviews, 78*(1), 1-39.

Den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D.-J., & Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of lipid research, 54*(9), 2325-2340.

Dijkshoorn, L., Ursing, B., & Ursing, J. (2000). Strain, clone and species: comments on three basic concepts of bacteriology. *Journal of medical microbiology, 49*(5), 397-401.

Du, Z., Hudcovic, T., Mrazek, J., Kozakova, H., Srutkova, D., Schwarzer, M., . . . Kverka, M. (2015). Development of gut inflammation in mice colonized with mucosa-associated bacteria from patients with ulcerative colitis. *Gut pathogens, 7*(1), 32.

Dubin, K., Callahan, M. K., Ren, B., Khanin, R., Viale, A., Ling, L., . . . Huttenhower, C. (2016). Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. *Nature communications, 7*(1), 1-8.

Eisen, J. A. (2007). Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol, 5*(3), e82.

Elinav, E., Strowig, T., Kau, A. L., Henao-Mejia, J., Thaiss, C. A., Booth, C. J., . . . Gordon, J. I. (2011). NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell, 145*(5), 745-757.

Everett, K. D., Bush, R. M., & Andersen, A. A. (1999). Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. *International Journal of Systematic and Evolutionary Microbiology, 49*(2), 415-440.

Eyre, D. W., Cule, M. L., Griffiths, D., Crook, D. W., Peto, T. E., Walker, A. S., & Wilson, D. J. (2013). Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission. *PLoS Comput Biol, 9*(5), e1003059.

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME journal, 6*(5), 1007-1017.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., . . . Van den Berghe, A. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature, 260*(5551), 500-507.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., . . . Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science, 269*(5223), 496-512.

Foster, J. T., Bull, R. L., & Keim, P. (2020). Ricin forensics: comparisons to microbial forensics. In *Microbial Forensics* (pp. 241-250): Elsevier.

Fournier, P.-E., Zhu, Y., Ogata, H., & Raoult, D. (2004). Use of highly variable intergenic spacer sequences for multispacer typing of Rickettsia conorii strains. *Journal of clinical microbiology, 42*(12), 5757-5766.

Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedeker, E. C., Harpaz, N., & Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences, 104*(34), 13780-13785.

Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J., & Huttenhower, C. (2015). Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences, 112*(22), E2930-E2938.

Fraser-Liggett, C. M. (2005). Insights on biology and evolution from microbial genome sequencing. *Genome Res, 15*(12), 1603-1610.

Garrett, W. S., Lord, G. M., Punit, S., Lugo-Villarino, G., Mazmanian, S. K., Ito, S., . . . Glimcher, L. H. (2007). Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. *Cell, 131*(1), 33-45.

Gensollen, T., Iyer, S. S., Kasper, D. L., & Blumberg, R. S. (2016). How colonization by microbiota in early life shapes the immune system. *Science, 352*(6285), 539-544.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., . . . Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science, 312*(5778), 1355-1359.

Glenn, T. C. (2011). Field guide to next ‐ generation DNA sequencers. *Molecular ecology resources, 11*(5), 759-769.

Greenblum, S., Turnbaugh, P. J., & Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences, 109*(2), 594-599.

Gueimonde, M., Ouwehand, A., Huhtinen, H., Salminen, E., & Salminen, S. (2007). Qualitative and quantitative analyses of the bifidobacterial microbiota in the colonic mucosa of patients with colorectal cancer, diverticulitis and inflammatory bowel disease. *World journal of gastroenterology: WJG, 13*(29), 3985.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology, 5*(10), R245-R249.

Heimesaat, M. M., Bereswill, S., Fischer, A., Fuchs, D., Struck, D., Niebergall, J., . . . Gescher, D. M. (2006). Gram-negative bacteria aggravate murine small intestinal Th1-type

immunopathology following oral infection with Toxoplasma gondii. *The Journal of Immunology, 177*(12), 8785-8795.

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., . . . Johnson, W. E. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome, 2*(1), 33.

Hooper, L. V., Midtvedt, T., & Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual review of nutrition, 22*(1), 283-307.

Huang, A., Kantor, R., DeLong, A., Schreier, L., & Istrail, S. (2011). QColors: an algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. *In silico biology, 11*(5, 6), 193-201.

Huber, J. A., Welch, D. B. M., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A., & Sogin, M. L. (2007). Microbial population structures in the deep marine biosphere. *Science, 318*(5847), 97-100.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol, 3*(2), reviews0003. 0001.

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology, 180*(18), 4765-4774.

Hugenholtz, P., & Tyson, G. W. (2008). Metagenomics. *Nature, 455*(7212), 481-483.

Hugon, P., Dufour, J.-C., Colson, P., Fournier, P.-E., Sallah, K., & Raoult, D. (2015). A comprehensive repertoire of prokaryotic species identified in human beings. *The Lancet Infectious Diseases, 15*(10), 1211-1219.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res, 17*(3), 377-386.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., . . . Madsen, K. L. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology, 7*, 459.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences, 109*(40), 16213-16216.

Kaput, J., Cotton, R. G., Hardman, L., Watson, M., Al Aqeel, A. I., Al‐Aama, J. Y., . . . Auerbach, A. D. (2009). Planning the human variome project: the Spain report. *Human mutation, 30*(4), 496-510.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664.

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res, 26*(12), 1721-1729.

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., . . . Stoye, J. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res, 36*(7), 2230-2239.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews, 72*(4), 557-578.

Lang, G. I., Botstein, D., & Desai, M. M. (2011). Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics, 188*(3), 647-661.

Lefébure, T., & Stanhope, M. J. (2007). Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol, 8*(5), R71.
Lepage, P., Häsler, R., Spehlmann, M. E., Rehman, A., Zvirbliene, A., Begun, A., . . . Raedler, A. (2011). Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology, 141*(1), 227-236.
Leung, H. C., Yiu, S.-M., Yang, B., Peng, Y., Wang, Y., Liu, Z., . . . Chin, F. Y. (2011). A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics, 27*(11), 1489-1495.
Levin, B. R., & Bergstrom, C. T. (2000). Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. *Proceedings of the National Academy of Sciences, 97*(13), 6981-6985.
Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell, 124*(4), 837-848.
Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.
Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., . . . Nielsen, T. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol, 32*(8), 834-841.
Li, W., Raoult, D., & Fournier, P.-E. (2009). Bacterial strain typing in the genomic era. *FEMS microbiology reviews, 33*(5), 892-916.
Li, X., Hu, H., & Li, X. (2020). mixtureS: a novel tool for bacterial strain reconstruction from reads. *Bioinformatics*.
Li, X., Naser, S. A., Khaled, A., Hu, H., & Li, X. (2018). When old metagenomic data meet newly sequenced genomes, a case study. *PLoS One, 13*(6), e0198773.
Li, X., Saadat, S., Hu, H., & Li, X. (2019). BHap: a novel approach for bacterial haplotype reconstruction. *Bioinformatics, 35*(22), 4624-4631.
Li, X., & Waterman, M. S. (2003). Estimating the repeat structure and length of DNA sequences using ℓ-tuples. *Genome Res, 13*(8), 1916-1922.
Lougheed, K. (2012). There are fewer microbes out there than you think. *Nature. doi, 10*, 13.
Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., & Gevers, D. (2015). ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol, 33*(10), 1045-1052.
Manichanh, C., Chapple, C. E., Frangeul, L., Gloux, K., Guigo, R., & Dore, J. (2008). A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res, 36*(16), 5180-5188.
Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380.
Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., . . . Anderson, I. (2007). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res, 36*(suppl_1), D534-D538.
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal, 17*(1), 10-12.
McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods, 4*(1), 63-72.

Monzoorul Haque, M., Ghosh, T. S., Komanduri, D., & Mande, S. S. (2009). SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics, 25*(14), 1722-1730.

Morris, R. M., Rappé, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A., & Giovannoni, S. J. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature, 420*(6917), 806-810.

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., . . . Gordon, J. I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science, 332*(6032), 970-974.

Nagy-Szakal, D., Hollister, E. B., Luna, R. A., Szigeti, R., Tatevian, N., Smith, C. W., . . . Kellermayer, R. (2013). Cellulose supplementation early in life ameliorates colitis in adult mice. *PLoS One, 8*(2), e56685.

Nakagawa, T., Ishibashi, J.-I., Maruyama, A., Yamanaka, T., Morimoto, Y., Kimura, H., . . . Fukui, M. (2004). Analysis of dissimilatory sulfite reductase and 16S rRNA gene fragments from deep-sea hydrothermal sites of the Suiyo Seamount, Izu-Bonin Arc, Western Pacific. *Applied and environmental microbiology, 70*(1), 393-403.

Nakanishi, Y., Sato, T., & Ohteki, T. (2015). Commensal Gram-positive bacteria initiates colitis by inducing monocyte/macrophage mobilization. *Mucosal immunology, 8*(1), 152-160.

Nayfach, S., Rodriguez-Mueller, B., Garud, N., & Pollard, K. S. (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res, 26*(11), 1612-1625.

Oh, J., Byrd, A. L., Deming, C., Conlan, S., Barnabas, B., Blakesley, R., . . . Dekhtyar, M. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature, 514*(7520), 59-64.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. Science 276: 734–740.

Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D., Buigues, B., . . . Auch, A. (2006). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science, 311*(5759), 392-394.

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One, 9*(4), e93827.

Prosperi, M. C., & Salemi, M. (2012). QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics, 28*(1), 132-133.

Pulido-Tamayo, S., Sánchez-Rodríguez, A., Swings, T., Van den Bergh, B., Dubey, A., Steenackers, H., . . . Marchal, K. (2015). Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. *Nucleic Acids Res, 43*(16), e105-e105.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . Yamada, T. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*(7285), 59-65.

Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G., & Eren, A. M. (2017). DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol, 18*(1), 1-22.

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Reviews in Microbiology, 57*(1), 369-394.

Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res, 38*(20), e191-e191.

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., . . . Minor, C. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and environmental microbiology, 66*(6), 2541-2547.

Rooks, M. G., Veiga, P., Wardwell-Scott, L. H., Tickle, T., Segata, N., Michaud, M., . . . Ballal, S. A. (2014). Gut microbiome composition and function in experimental colitis during active disease and treatment-induced remission. *The ISME journal, 8*(7), 1403-1417.

Roosaare, M., Vaher, M., Kaplinski, L., Möls, M., Andreson, R., Lepamets, M., . . . Remm, M. (2017). StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ, 5*, e3353.

Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics, 27*(1), 127-129.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., . . . Remington, K. (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol, 5*(3), e77.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., . . . Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature, 265*(5596), 687-695.

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual review of microbiology, 31*(1), 107-133.

Schirmer, M. (2014). *Algorithms for viral haplotype reconstruction and bacterial metagenomics: resolving fine-scale variation in next generation sequencing data.* University of Glasgow,

Schroeder, B. O., & Bäckhed, F. (2016). Signals from the gut microbiota to distant organs in physiology and disease. *Nature medicine, 22*(10), 1079.

Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications, 4*(1), 1-11.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods, 9*(8), 811-814.

Simon, C., & Daniel, R. (2011). Metagenomic analyses: past and future trends. *Applied and environmental microbiology, 77*(4), 1153-1161.

Smillie, C. S., Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., . . . Sadowsky, M. J. (2018). Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell host & microbe, 23*(2), 229-240. e225.

Sobkowiak, B., Glynn, J. R., Houben, R. M., Mallard, K., Phelan, J. E., Guerra-Assunção, J. A., . . . McNerney, R. (2018). Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. *BMC Genomics, 19*(1), 613.

Stackebrandt, E., & GOEBEL, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology, 44*(4), 846-849.

Stark, M., Berger, S. A., Stamatakis, A., & von Mering, C. (2010). MLTreeMap-accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics, 11*(1), 461.

Suau, A., Bonnet, R., Sutren, M., Godon, J.-J., Gibson, G. R., Collins, M. D., & Doré, J. (1999). Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and environmental microbiology, 65*(11), 4799-4807.

Surget-Groba, Y., & Montoya-Burgos, J. I. (2010). Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res, 20*(10), 1432-1440.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., . . . Durkin, A. S. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences, 102*(39), 13950-13955.

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal, 474*(11), 1823-1836.

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res, 27*(4), 626-638.

Turnbaugh, P. J., & Gordon, J. I. (2008). An invitation to the marriage of metagenomics and metabolomics. *Cell, 134*(5), 708-713.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature, 449*(7164), 804-810.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., . . . Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature, 428*(6978), 37-43.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., . . . Nelson, W. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74.

Vigsnæs, L. K., Brynskov, J., Steenholdt, C., Wilcks, A., & Licht, T. R. (2012). Gram-negative bacteria account for main differences between faecal microbiota from patients with ulcerative colitis and healthy controls. *Beneficial microbes, 3*(4), 287-297.

Wang, Y. (2016). *Computational Approaches for Binning Metagenomic Reads.* University of Central Florida,

Wang, Y., Hu, H., & Li, X. (2015). MBBC: an efficient approach for metagenomic binning based on clustering. *BMC bioinformatics, 16*(1), 36.

Wang, Y., Hu, H., & Li, X. (2016). MBMC: An effective Markov chain approach for binning metagenomic reads from environmental shotgun sequencing projects. *Omics: a journal of integrative biology, 20*(8), 470-479.

Wang, Y., Hu, H., & Li, X. (2017). rRNAFilter: A Fast Approach for Ribosomal RNA Read Removal Without a Reference Database. *Journal of Computational Biology, 24*(4), 368-375.

Wang, Y., Leung, H. C., Yiu, S.-M., & Chin, F. Y. (2012). MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics, 28*(18), i356-i362.

Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M., . . . Stackebrandt, E. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial

systematics. *International Journal of Systematic and Evolutionary Microbiology, 37*(4), 463-464.
Wayne, L. G. (1988). International Committee on Systematic Bacteriology: announcement of the report of the ad hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Zentralblatt fur Bakteriologie, Mikrobiologie, und Hygiene. Series A, Medical microbiology, infectious diseases, virology, parasitology, 268*(4), 433.
Wexler, H. M. (2007). Bacteroides: the good, the bad, and the nitty-gritty. *Clinical microbiology reviews, 20*(4), 593-621.
Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol, 15*(3), 1-12.
Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput Biol, 6*(2), e1000667.
Wu, Y.-W., & Ye, Y. (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology, 18*(3), 523-534.
Xu, K., & Jiang, B. (2017). Analysis of mucosa-associated microbiota in colorectal cancer. *Medical science monitor: international medical journal of experimental and clinical research, 23*, 4422.
Ye, J., Lee, J. W., Presley, L. L., Bent, E., Wei, B., Braun, J., . . . Borneman, J. (2008). Bacteria and bacterial rRNA genes associated with the development of colitis in IL-10−/− mice. *Inflammatory bowel diseases, 14*(8), 1041-1050.
Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., . . . Li, W. (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol, 5*(3), e16.
Zagordi, O., Bhattacharya, A., Eriksson, N., & Beerenwinkel, N. (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC bioinformatics, 12*(1), 1-5.
Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res, 18*(5), 821-829.
Zinner, S. H. (1999). Changing epidemiology of infections in patients with neutropenia and cancer: emphasis on gram-positive and resistant bacteria. *Clinical infectious diseases, 29*(3), 490-494.
Zolfo, M., Tett, A., Jousson, O., Donati, C., & Segata, N. (2017). MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res, 45*(2), e7-e7.